

Efficient real-time monitoring of an emerging influenza epidemic: how feasible?

Paul J. Birrell¹, Daniela De Angelis^{1,2}, Lorenz Wernsich¹, Brian D. M. Tom¹,
Gareth O. Roberts³, Richard G. Pebody²

August 19, 2016

¹*MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way,
Cambridge CB2 0SR, UK*

²*Public Health England, London, UK*

³*Centre for Research in Statistical Methodology, University of Warwick, Coventry, UK*
e-mail for correspondence: daniela.deangelis@mrc-bsu.cam.ac.uk

Abstract

A prompt public health response to a new epidemic relies on the ability to monitor and predict its evolution in real time as data accumulate. The 2009 A/H1N1 outbreak in the UK revealed pandemic data as noisy, contaminated, potentially biased, and originating from multiple sources, seriously questioning the capacity for real-time monitoring. Here we assess the feasibility of real-time inference based on such data by constructing an analytic tool combining an age-stratified SEIR transmission model with various observation models describing the data generation mechanisms. As batches of data become available, a sequential Monte Carlo algorithm is developed to synthesise multiple imperfect data streams and iterate epidemic inferences amidst rapidly evolving epidemic environments, heuristically minimising computation time to ensure timely delivery of real-time epidemic assessments.

KEYWORDS: Sequential Monte-Carlo, Resample-Move, real-time inference, pandemic influenza, SEIR transmission model

1 Introduction

A pandemic influenza outbreak has the potential to place a significant burden upon healthcare systems. Therefore, the capacity to monitor and predict the evolution of an epidemic as data progressively accumulate is a key component of preparedness strategies for prompt public health response.

Statistical inferential approaches have been used in a real-time monitoring context for a number of infectious diseases. Examples include: prediction of swine fever cases in a classical framework Meester et al. (2002); online estimation of a time-evolving effective reproduction number $R(t)$ for SARS (Wallinga and Teunis, 2004; Cauchemez et al., 2006) and for a generic emerging disease (Bettencourt and Ribeiro, 2008); and Bayesian inference on the transmission dynamics of avian influenza in the UK poultry industry (Jewell et al., 2009).

These models rely on the availability of direct data on the number of new cases of an infectious disease over time. In practice, as illustrated by the 2009 outbreak of pandemic A/H1N1pdm influenza in the United Kingdom (UK), direct data are seldom available. More likely, multiple sources of data exist, each indirectly informing the epidemic evolution, each subject to possible sources of bias. This calls for more complex modelling, requiring the synthesis of information from a range of data sources in real time.

In this paper we tackle the problem of online inference on an influenza pandemic in this more realistic situation. To address this problem we develop the work of Birrell et al. (2011) who retrospectively reconstructed the A/H1N1 pandemic in a Bayesian framework using multiple data streams collected over the course of the pandemic. In Birrell et al. (2011) posterior distributions of relevant epidemic parameters and related quantities are derived through Markov Chain Monte Carlo (MCMC) methods which, if used in real-time, pose important computational challenges. MCMC is notoriously inefficient for online inference as it requires repeat browsing of the full history of the data as new data accrues. This motivates a more efficient algorithm. Potential alternatives include refinements of MCMC (e.g. Jewell et al., 2009; Banterle et al., 2015) and Bayesian emulation as in Farah et al. (2014), where the model is replaced by an easily-evaluated approximation that can be readily prepared in advance of the data assimilation process. Here, we explore Sequential Monte Carlo (SMC) methods (Doucet and Johansen, 2009) as an alternative to the expensive MCMC simulations. As batches of data arrive at times t_1, \dots, t_K , SMC techniques allow computationally efficient online inference by combining the posterior distribution $\pi_k(\cdot)$ at time t_k with the incoming batch of data to obtain an estimate for $\pi_{k+1}(\cdot)$.

Use of SMC in the real time monitoring of an emerging epidemic is not new. Dureau et al. (2013), Camacho et al. (2015), Dukic et al. (2012), Ong et al. (2010), and Skvortsov and Ristic (2012) are examples of real time estimation and prediction for deterministic and stochastic epidemic systems describing the dynamics of influenza and Ebola epidemics. Their models, however, also only include a single source of information that has either been pre-smoothed or is free of any sudden or systematic changes.

In what follows we advance existing literature in two ways: we include a number of data streams, realistically mimicking the 2009 pandemic in the UK; and we consider the situation where a public health intervention introduces a shock to the system, critically disrupting the ability to track the posterior distribution over time.

The paper is organised as follows: in Section 2 the model in Birrell et al. (2011) is reviewed focusing on the data available and the computational limitations of the MCMC algorithm in a real time context; in Section 3 the idea of SMC is introduced and an algorithm based on the work in Gilks and Berzuini (2001) described; in Sections 4 and 5 results are presented from the application of a naive SMC algorithm to data simulated to mimic the 2009 outbreak and illustrate the challenges posed by the presence of the informative observations induced by system shocks; in Sections 6 and 7 adapted SMC approaches that address such challenges are assessed; we conclude with Section 8 in which the ideas explored in the paper are critically reviewed and outstanding issues discussed.

2 A Model For Pandemic Reconstruction

Birrell et al. (2011) describe the transmission of a novel influenza virus among a fixed population stratified into A age groups and the subsequent reporting of infections. This is achieved through using a deterministic age-structured Susceptible (S), Exposed (E), Infectious (I), Recovered (R) transmission model, with the E and I states split into two sub-states, E_1 and E_2, I_1

and I_2 . At time $t_k = k\delta t$ ($k = 0, \dots, K$) the vector $(S_{t_k,a}, E_{1t_k,a}, E_{2t_k,a}, I_{1t_k,a}, I_{2t_k,a})$ gives the number of individuals in age group a ($a = 1, \dots, A$) in each model state. The dynamics of the system are governed by a set of difference equations, such that for suitably small increments δt :

$$\begin{aligned} S_{t_k,a} &= S_{t_{k-1},a} (1 - \lambda_{t_{k-1},a} \delta t) \\ E_{1t_k,a} &= E_{1t_{k-1},a} \left(1 - \frac{2\delta t}{d_L}\right) + S_{t_{k-1},a} \lambda_{t_{k-1},a} \delta t \\ E_{2t_k,a} &= E_{2t_{k-1},a} \left(1 - \frac{2\delta t}{d_L}\right) + E_{1t_{k-1},a} \frac{2\delta t}{d_L} \\ I_{1t_k,a} &= I_{1t_{k-1},a} \left(1 - \frac{2\delta t}{d_I}\right) + E_{2t_{k-1},a} \frac{2\delta t}{d_L} \\ I_{2t_k,a} &= I_{2t_{k-1},a} \left(1 - \frac{2\delta t}{d_I}\right) + I_{1t_{k-1},a} \frac{2\delta t}{d_I} \end{aligned} \quad (1)$$

where d_L and d_I are the mean latent and the mean infectious periods respectively. Transmission is driven by the time- and age-varying force of infection, $\lambda_{t,a}$, the rate at which susceptible individuals become infected:

$$\lambda_{t_k,a} = 1 - \prod_{b=1}^A \left\{ \left(1 - M_{t_k}^{*(a,b)} R_0(\psi) / d_I\right)^{I_{1t_k,b} + I_{2t_k,b}} \right\}. \quad (2)$$

Here, $R_0(\psi)$ is the basic reproduction number, the expected number of secondary infections caused by a single primary infection in a fully susceptible population, parameterised in terms of the epidemic growth rate ψ . The pattern of transmission between age groups is determined by time-varying mixing matrices M_{t_k} , with $M_{t_k}^{(a,b)}$ giving relative rates of effective contacts between individuals of each pair of age groups (a, b) . These matrices are scaled to have elements

$$M_{t_k}^{*(a,b)} = M_{t_k}^{(a,b)} / R_0^*$$

where R_0^* is the dominant eigenvalue of the time-0 next generation matrix whose $(a, b)^{\text{th}}$ entry is $M_{t_0}^{(a,b)} N_a$, with N_a being the population size of the a^{th} age stratum. The initial conditions of the system are determined by: parameter I_0 , the total number of infectious individuals across all age groups at time t_0 ; an assumed equilibrium distribution of infections over the age groups; and an assumption of initial exponential growth that determines the relationship between the numbers in the four disease states. For ease of implementation, a reparameterisation is made from I_0 to a parameter denoted v , the details of which can be found in the Supplementary Information to Birrell et al. (2011).

Denote by $\xi = (\psi, v, d_I, m)$ the vector of transmission dynamics parameters where m parameterise the mixing matrices M_{t_k} , defining any time variation. Note, parameter d_L is notoriously difficult to estimate and is therefore assumed fixed at two days.

2.1 Distributional assumptions

Figure 1 illustrates how surveillance data from multiple sources relate to the age-structured SEIR transmission model, allowing estimation of the transmission dynamics parameters. The transmission process is unobserved. However, there are a number of surveillance sources informing aspects of this process. As system dynamics are assumed to be deterministic, there is no system error and outputs (see Equations (1)-(2)) are deterministic functions of ξ , e.g.

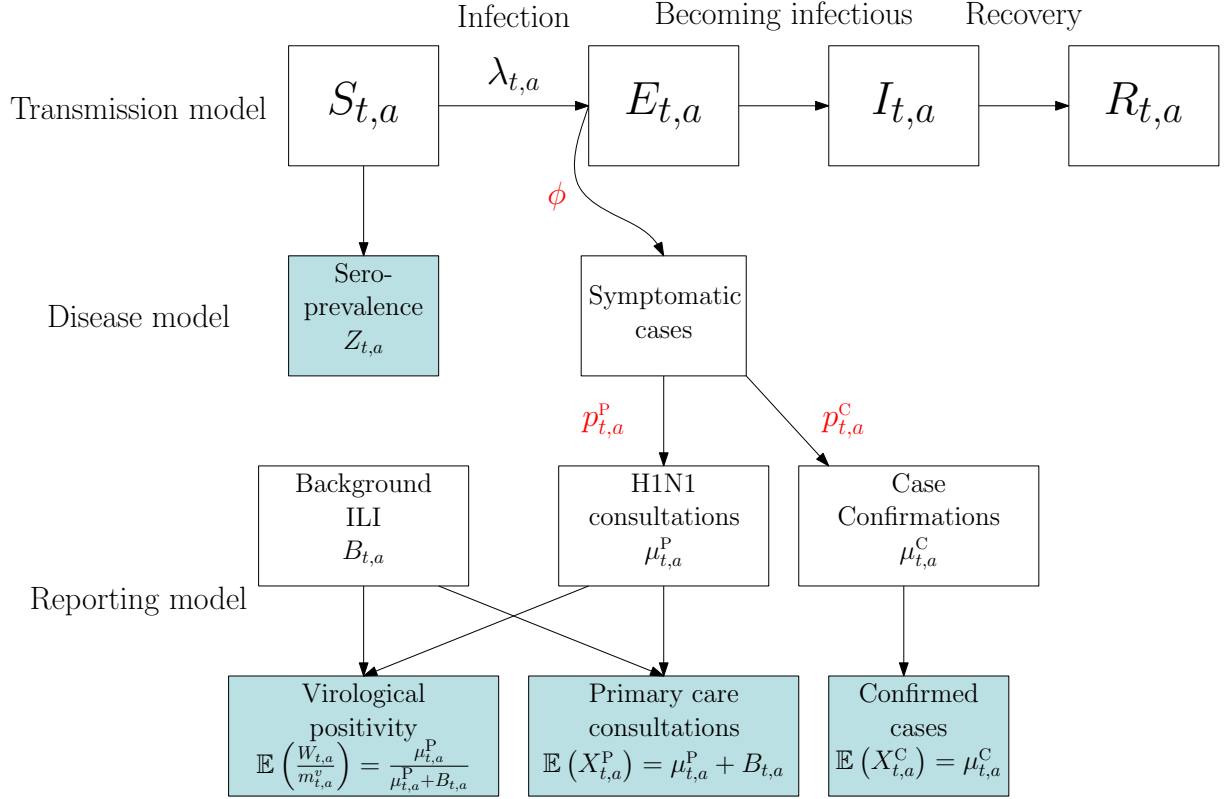


Figure 1: Schematic diagram of how multiple epidemics surveillance sources can link into an SEIR epidemic model via an observation and reporting model (Birrell et al., 2011). Shaded boxes represent observed data streams.

$S_{t_k,a} \equiv S_{t_k,a}(\xi)$. The available surveillance data are ‘imperfect observations on a perfect system’ and are linked to the system’s outputs via observational models as follows.

The number of susceptibles in age group a at the end of the k^{th} time-step, $S_{t_k,a}(\xi)$, is informed directly by a series of cross-sectional survey data $Z_{t_k,a}$ on the presence of immunity-conferring antibodies in the general population. Denoting by $m_{t_k,a}^s$ the number of blood sera samples tested in time interval $[t_{k-1}, t_k)$, it is assumed that

$$Z_{t_k,a} \sim \text{Bin} \left\{ m_{t_k,a}^s, 1 - \frac{S_{t_k,a}(\xi)}{N_a} \right\}. \quad (3)$$

The number of new age-specific infections in interval $[t_{k-1}, t_k)$ expressed as

$$\Delta_{t_k,a}(\xi) = S_{t_{k-1},a}(\xi) \lambda_{t_{k-1},a}(\xi) \delta t.$$

are indirectly related to surveillance data on health-care burden. A proportion, ϕ (see Figure 1) of these new infections will develop symptoms. Of those symptomatic, a proportion $p_{t_k,a}^C$ will be virologically confirmed through admission to hospital and/or to an intensive care unit (ICU). Alternatively a proportion $p_{t_k,a}^P$ will choose to contact primary care practitioners and will be reported as consultations for influenza-like illness (ILI) together with individuals attending for non-pandemic pathogen ILI. As a result, primary consultation data will be contaminated by a background consultation component strongly influenced by the public’s volatile sensitivity to governmental advice. The consultation data are, therefore, less directly related to the severity and incidence of infection than the confirmed cases. To identify the consultations attributable

to the pandemic strain, complementary data from a sub-sample of swabbed ILI patients provide information on the proportion of consultations with pandemic virus.

Using a generic $e \in (C, P)$ to denote counts of confirmed cases or primary care consultations, the model in Birrell et al. (2011) outputs quantities of the type

$$\mu_{t_k,a}^e = \phi p_{t_k,a}^e \sum_{l=0}^k \Delta_{t_{k-l},a} f_{\zeta_e, \sigma_e^2}(l), \quad (4)$$

representing the number of surveillance counts in the interval $[t_{k-1}, t_k)$ attributable to the pandemic. Expression (4) results from the process of becoming infected and subsequently experiencing a delay (of mean ζ_e , variance σ_e^2 with discretised probability mass function $f_{\zeta_e, \sigma_e^2}(\cdot)$), which includes the time from infection to symptoms (the incubation period), the time from symptoms to the healthcare event, and the time from diagnosis to the report of the healthcare event of interest. Note that in (4) the parametric dependence of output quantities has been omitted for ease of notation and will be done throughout.

The count data $X_{t_k,a}^e$ are assumed to have negative binomial distribution expressed here in mean-dispersion (μ, η) parameterisation, such that if $X \sim \text{NegBin}(\mu, \eta)$, then $\mathbb{E}(X) = \mu$, $\text{var}(X) = \mu(\eta + 1)$ and

$$\mathbb{P}(X = x) = \binom{x + \frac{\mu}{\eta} - 1}{x} \left(\frac{1}{1 + \eta} \right)^{\frac{\mu}{\eta}} \left(\frac{\eta}{1 + \eta} \right)^x. \quad (5)$$

So, for the confirmed cases $X_{t_k,a}^C$:

$$X_{t_k,a}^C \sim \text{NegBin}(\mu_{t_k,a}^C, \eta_{t_k}) \quad (6)$$

and for the primary care consultations $X_{t_k,a}^P$, which include contamination by a non-pandemic ILI background component $B_{t_k,a}$:

$$X_{t_k,a}^P \sim \text{NegBin}(\mu_{t_k,a}^P + B_{t_k,a}, \eta_{t_k}) \quad (7)$$

where the contamination $B_{t_k,a}$ is appropriately parameterised in terms of parameters β^B , and the signal $\mu_{t_k,a}^P$ is identified by virological data from sub-samples of size $m_{t_k,a}^v$ of the primary care consultations. The number of swabs testing positive for the presence of the pandemic strain $W_{t_k,a}$ in each sample is assumed to be distributed as:

$$W_{t_k,a} \sim \text{Bin} \left(m_{t_k,a}^v, 1 - \frac{B_{t_k,a}}{\mu_{t_k,a}^P + B_{t_k,a}} \right) \quad (8)$$

2.2 Implementation

Let θ denote the vector of all free parameters *i.e.* $\theta = \{\xi, \phi, p_{t_k,a}^e, \eta_{t_k}, \beta^B\}$. Birrell et al. (2011) develop a Bayesian approach and use a Markov Chain-Monte Carlo (MCMC) algorithm to derive the posterior distribution of θ on the basis of 245 days of primary care consultation and swab positivity data, confirmed case and cross-sectional serological data.

The MCMC algorithm is a naively adaptive random walk Metropolis algorithm, requiring 7×10^5 iterations, taking over four hours. MCMC is not easily adapted for parallelised computation, although a small speed up can be achieved by parallelising the likelihood component of the posterior distribution of θ over a small number of CPUs. In total, this required in excess of

6.3×10^6 evaluations of the transmission model and/or convolutions of the kind in equation (4). Implementation of MCMC in an online fashion, as new data arrive involves the re-analysis of the entire dataset, requiring time for multiple Markov chains to converge.

Although, the runtime might not be prohibitive for real-time inference, the current implementation leaves little margin to consider multiple code runs or alternative model formulations. In a future pandemic there will be a greater wealth of data facilitating a greater degree of stratification of the population (Scientific Pandemic Influenza Advisory Committee (SPI): Subgroup on Modelling, 2011). With increasing model complexity comes rapidly increasing MCMC run-times, which can be efficiently addressed through use of SMC methods.

3 An SMC Alternative to MCMC

SMC is commonly used for inference from models that can be cast in a state-space formulation, where expressions of the form:

$$\begin{aligned} X_k &\sim p(\cdot | X_{k-1} = x_{k-1}, \theta) \\ Y_k &\sim q(\cdot | X_k = x_k, \theta) \end{aligned} \quad (9)$$

govern the evolution of the latent state vector x_k and the its relation to the observed data y_k for $k = 1, \dots, K$. Here the Y_k are conditionally independent given knowledge of the x_k . On observing k batches of data, y_1, \dots, y_k at times t_1, \dots, t_k , the main interest in this set-up is the filtering problem i.e. estimating the state vector x_k through posterior distributions $\pi_k(x_k | y_{1:k}, \theta)$. Note the conditioning on the parameters θ , which are typically assumed to be fixed and known (Cappé et al., 2007), although methods for the estimation of static θ are very much an active area of research (for example Martin, 2012).

The model in Section 2 is a deterministic model, designed for use at a time in a pandemic when stochastic effects are uninfluential. In this case $x_k \equiv x_k(\theta)$ with data being imperfect observations distributed around model outputs. The inferential focus is here on θ . On-line inference involves the sequential estimation of posterior distributions $\pi_k(\theta) = p(\theta | y_{1:k}) \propto \pi_0(\theta) p(y_{1:k} | \theta)$, $k = 1, \dots, K$, where $\pi_0(\theta)$ indicates the prior for θ . Estimation of any epidemic feature, *e.g.* the assessment of the current state of the epidemic or prediction of its future course, follows from estimating θ (or components thereof).

Suppose at time t_k a set of particles $\{\theta_k^{(1)}, \dots, \theta_k^{(n_k)}\}$, where each particle $\theta_k^{(j)}$ carries a weight $\omega_k^{(j)}$, approximates a sample from the target distribution $\pi_k(\cdot)$. On the arrival of the next batch of data, $\pi_k(\cdot)$ is then used as an importance sampling distribution to sample from $\pi_{k+1}(\cdot)$. In practice, this involves a re-weighting of the particle set. From the conditional independence assumption of (9), the particles are reweighted according to the importance ratio:

$$\omega_{k+1}^{(j)} \propto \omega_k^{(j)} \frac{\pi_{k+1}(\theta_k^{(j)})}{\pi_k(\theta_k^{(j)})} = \omega_k^{(j)} p(y_{k+1} | \theta_k^{(j)});$$

which reduces to the likelihood of the incoming data batch. Eventually, many particles will begin to carry relatively very low weight, leading to sample degeneracy as progressively fewer particles contribute meaningfully to the estimation of $\pi_k(\cdot)$. A measure of this degeneracy is the effective sample size (ESS) (Liu and Chen, 1995),

$$\text{ESS}(\{\omega_k^{(\cdot)}\}) = \frac{\left(\sum_{l=1}^{n_k} \omega_k^{(l)}\right)^2}{\sum_{j=1}^{n_k} \omega_k^{(j)2}}. \quad (10)$$

Values for the ESS that are small in comparison to n_k are indicative of degeneracy or impoverishment of the current particle set.

This degeneracy can be tackled in different ways. Gordon et al. (1993) introduced a resampling step, removing low weight particles and re-setting particle weights, and proposed jittering the particles. This jittering step was later formalised by Gilks and Berzuini (2001) with the introduction of Metropolis-Hastings (MH) steps to rejuvenate the sample. Fearnhead (2002) and Chopin (2002) provide more general treatises of this sequential Monte Carlo method, with Chopin (2002) labelling the algorithm ‘iterated batch importance sampling’. This approach has since been extended by Del Moral et al. (2006) who unify the static estimation with the filtering problem (estimation of x_k).

Here we adapt the resample-move algorithm of Gilks and Berzuini (2001) and investigate its potential efficiency saving when compared to successive use of MCMC. MH steps provide the computational bottle-neck in resample-move as they require the browsing of the whole history of the data to evaluate the full likelihood, not just the latest batch of observations. To achieve fast inference, it is preferable to limit the number of such steps, without introducing Monte Carlo error through having a degenerate sample.

The algorithm is laid out in full below. It is presumed that it is straightforward to sample prior distribution $\pi_0(\theta)$.

3.1 The Algorithm

1. **Set** $k = 0$. Draw a sample $\{\theta_0^{(1)}, \dots, \theta_0^{(n_0)}\}$ from the prior distribution, $\pi_0(\theta)$, set the weights $\omega_0^{(j)} = 1/n_0, \forall j$.
2. **Set** $k = k + 1$. Observe a new batch of data $Y_k = y_k$. Re-weighted the particles so that the j^{th} particle now has weight

$$\tilde{\omega}_k^{(j)} \propto \omega_{k-1}^{(j)} p(y_k | \theta_{k-1}^{(j)}). \quad (11)$$

3. **Calculate the effective sample size.** Set $\omega_k^{*(j)} = \tilde{\omega}_k^{(j)} / \sum_i \tilde{\omega}_k^{(i)}, \forall j$. If $ESS\left(\left\{\omega_k^{*(\cdot)}\right\}\right) > \varepsilon_L n_{k-1}$ set $\theta_k^{(j)} = \theta_{k-1}^{(j)}, \omega_k^{(j)} = \omega_k^{*(j)}, n_k = n_{k-1}$ and return to point (2), else go next.
4. **Resample.** Choose n_k and sample $\{\tilde{\theta}_k^{(j)}\}_{j=1}^{n_k}$ from the set of particles $\{\theta_{k-1}^{(j)}\}_{j=1}^{n_{k-1}}$ with corresponding probabilities $\{\omega_k^{*(j)}\}_{j=1}^{n_{k-1}}$. Here, we have used residual resampling (Liu and Chen, 1998). Re-set $\omega_k^{(j)} = 1/n_k$.
5. **Move:** For each j , move from $\tilde{\theta}_k^{(j)}$ to $\theta_k^{(j)}$ via a MH kernel $\mathcal{K}_k(\tilde{\theta}_k^{(j)}, \theta_k^{(j)}; \gamma)$. If $k < K$, return to point (2).

6. **End.**

There are a number of algorithmic choices to be made, including tuning the parameters of the MH kernel (γ above) or the rejuvenation threshold, ε_L . In a real-time setting, it may not be possible to tune an algorithm “on the fly”, so the system has to be able to work “out of the box”, either through prior tuning or through adaptation. In what follows we set $\varepsilon_L = 0.5$ (Jasra et al., 2011) and we focus on the key factor affecting the performance of the algorithm in real-time, *i.e.* the MH kernel.

3.1.1 Kernel Choice

Correlated Random Walk A correlated random walk proposes values:

$$\theta^* | \tilde{\theta}_k^{(j)} \sim N(\tilde{\theta}_k^{(j)}, \gamma \bar{\Sigma}_k) \quad (12)$$

in the neighbourhood of the current particle, where $\bar{\Sigma}_k$ is the sample variance-covariance matrix for the weighted sample $\{\tilde{\omega}_k^{(\cdot)}, \theta_{k-1}^{(\cdot)}\}$. The parameter γ can be tuned *a priori* to guarantee a reasonable acceptance rate, or, alternatively, asymptotic results for the optimal scaling of covariance matrices (Roberts and Rosenthal, 2001; Sherlock et al., 2010) can be used. Localised moves keep acceptance rates high and will quickly restore the value of the ESS. However, if after re-sampling there are few unique particles then the rejuvenation will result in a highly clustered sample, providing an inaccurate representation of the target distribution.

Approximate Gibbs' An independence sampler that proposes (Chopin, 2002):

$$\theta^* | \tilde{\theta}_k^{(j)} \sim N(\bar{\theta}_k, \bar{\Sigma}_k) \quad (13)$$

where $\bar{\theta}_k$ is the sample mean for the $\{\tilde{\omega}_k^{(\cdot)}, \theta_{k-1}^{(\cdot)}\}$.

Here, moves are proposed to a region of the sample space only weakly dependent on the current position and proposals are drawn from a distribution chosen to approximate the target distribution. An accept-reject step is still required to correct for this approximation, so it is perhaps more accurate to refer to this proposal kernel as Approximate Metropolis-within-Gibbs'. The quality of the approximation depends on $\pi_{k-1}(\cdot)$ being well represented by the current particle set, there being sufficient richness in the particle weights after the re-weighting step and the target density being sufficiently near-Gaussian. Assuming that the multivariate normal approximation to the target is adequate (and it should be increasingly so as more data are acquired) this type of proposal allows for more rapid exploration of the sample space. If the multivariate normal approximation is not good, particles of high posterior or low proposal density will not be easily moved by this kernel, and, as acceptance rates cannot be adapted to ensure a minimum level of acceptance, there is no guarantee that the ESS will be restored above the level at which rejuvenations are required (see Section 5).

Both the correlated random walk and the approximate Gibbs' methods will be used, both as block updates where a new value for the entire parameter vector is proposed at once, and component-wise updates where individual or small groups of parameter components are proposed in turn, using the appropriate conditional distributions derived from (12) and (13).

4 A simulated epidemic

The SMC algorithm's performance against the gold-standard MCMC is evaluated via simulation, through its application to data arising from an epidemic simulated to mimic the timing and dynamics of the 2009 A/H1N1 pandemic in England.

Anomalously, this epidemic started with an initial burst of infection in Spring, so we assume that the starting date is the 1st May. The epidemic occurs in two waves of infection, the first reaches a peak immediately prior to the summer school holidays. After the holiday, the growth of the epidemic is far slower, reaching a second peak in the Autumn.

We consider two scenarios. In the first scenario we have direct information on confirmed cases, as might arrive in the surveillance of severe disease (e.g. hospitalisation, ICU admissions). In the second scenario we observe ILI consultations in primary care which are noisy

Table 1: Parameters used in the simulation of (confirmed case) epidemic data

<i>Parameter</i>	<i>Description</i>	<i>Value</i>
η	Dispersion parameters, for the primary care consultation data. Split either side of a public health intervention at $t_k = 83$, denoted (η_1, η_2) . 2 parameters	(3.00, 2.15)
d_I	Parameter describing the average infectious period. 1 parameter	3.47
ϕ	Parameter describing the proportion symptomatic. 1 parameter	0.278
m	Multipliers applied to the contact matrices (e.g. to describe the school-holiday effects). 5 parameters	(0.403, 0.495, 0.0588, 0.301, 0.421)
ψ	Exponential growth rate. 1 parameter	0.133
v	A reparameterisation of the initial number of infectives, a function of I_0 . 1 parameter	-13.9
p^C or p^P	Parameters governing the population propensity of individuals with ILI symptoms to appear in the data. Split either side of the public health intervention at $t_k = 83$, with different rates for adults and children. 4 parameters	$p_{t_k, a}^e = \begin{cases} p_1 & t_k \leq 83, a < 4 \\ p_2 & t_k \leq 83, a \geq 4 \\ p_3 p_1 & t_k > 83, a < 4 \\ p_4 p_2 & t_k > 83, a \geq 4 \end{cases}$ $p = (0.278, 0.162, 0.137, 0.441)$

and contaminated by non-pandemic infections (see Section 2.1). Both confirmed case and consultation data are assumed to exist alongside serological data measuring the overall level of cumulative infection over the course of the pandemic to date. In the second scenario, we also assume the existence of a companion dataset of virological swabbing data from a sub-sample of the noisy data. In both scenarios, observations are assumed to be made on 245 consecutive days and the underlying epidemic curve is characterised by the same parameters, so both confirmed case and primary care consultation data are subject to similar trends and shocks.

One such shock arises from an assumed sudden change in the way case counts, whether they are confirmed cases or GP consultations, are reported. This could occur due to some public health intervention, as happened in 2009 with the launch of the National Pandemic Flu Service (NPFs), designed to alleviate the burden placed on primary care services. Table 1 presents the model parameters common to both scenarios and the values used for simulation.

4.1 Confirmed Case Data

For a given set of parameters θ , the number of confirmed cases, $\mu_{t_k, a}^C$, in interval $[t_{k-1}, t_k)$ is given by Equation (4). Count data $X_{t_k, a}^C$ are then generated as negative binomially distributed

with mean $\mu_{t_k,a}^C(\theta)$ and variance $(1 + \eta_{t_k})\mu_{t_k,a}^C(\theta)$. The degree of overdispersion, defined by the dispersion parameters η_1 and η_2 (Table 1), is piecewise constant over time, with a breakpoint at the time of the system shock, taken to be $t_k = 83$.

4.2 Primary Care Consultation Data

These data contain a significant amount of contamination. As with the confirmed case data, the number of consultations due to the pandemic strain is calculated via the convolution equation (4) to give $\mu_{t_k,a}^P(\theta)$. The contamination component is added by assuming ‘background’ consultation rates that follow a log-linear spline with a discontinuity at $t_k = 83$, with additional age effects to generate separate consultation rates for children (< 16 year-olds) and adults. The background rates over the interval $[t_{k-1}, t_k)$, $B_{t_k,a}$, depend on spline parameters β^B , such that, for a suitable design matrix H^B

$$B = \exp\left(H^B \beta^B\right)$$

where B is a suitably vectorised collection of the $B_{t_k,a}$. Aggregated over ages, the log-linear spline used for simulation is plotted in Figure 2(D). In this example, β^B is a 9-dimensional parameter.

As already anticipated above, these counts will also drop markedly due to an intervention to reduce the burden on primary care services, resulting in a sudden change in the parameter p^P , the proportion of symptomatic cases that seek consultation. In reality, this parameter will show more heterogeneity over time than its analog for the confirmed case data as it depends on behavioural factors and is not a property of the virus. However, in the examples presented here, $p_{t_k,a}^C$ and $p_{t_k,a}^P$ are parameterised similarly (see Table 1).

From (7), the expected number of observed consultations is $B_{t_k,a} + \mu_{t_k,a}^P(\theta)$. Again, a negative binomial distribution is used to generate the data, assuming the same parameterisation and levels of dispersion used in the generation of the confirmed cases. The companion dataset to the consultation counts is the virological swabbing data. On each day of the simulated epidemic, the number of swabs of patients taken is comparable to the number of swabs on the same day in 2009. These swabs are assumed to be representative of the consulting population and it is assumed the test is entirely reliable. These data are assumed to have binomial distribution, as laid out in (8).

4.3 Serological Data

For reliable real-time assessment of pandemic progression, the timely availability of serological data is vital. Serological data arise from the testing of blood sera samples, ideally taken representatively from the entire population. The proportion of positive tests over the course of the epidemic gives an indication of the level of cumulative incidence, effectively measuring the scale of the epidemic. Typically, these data are not available at the beginning of an epidemic outbreak, as the relevant assay for the circulating influenza strain might take some time to be developed. Again, the 2009 pandemic is taken as a guide, with sample sizes and the timings of samples being taken from this pandemic.

The seropositivity data, $\{Z_{t_k,a}\}$, are simulated from the binomial distribution given in Equation (3), with parameter values as in Table 1.

All the data for the second scenario (primary healthcare data plus serological data) are presented in Figure 2.

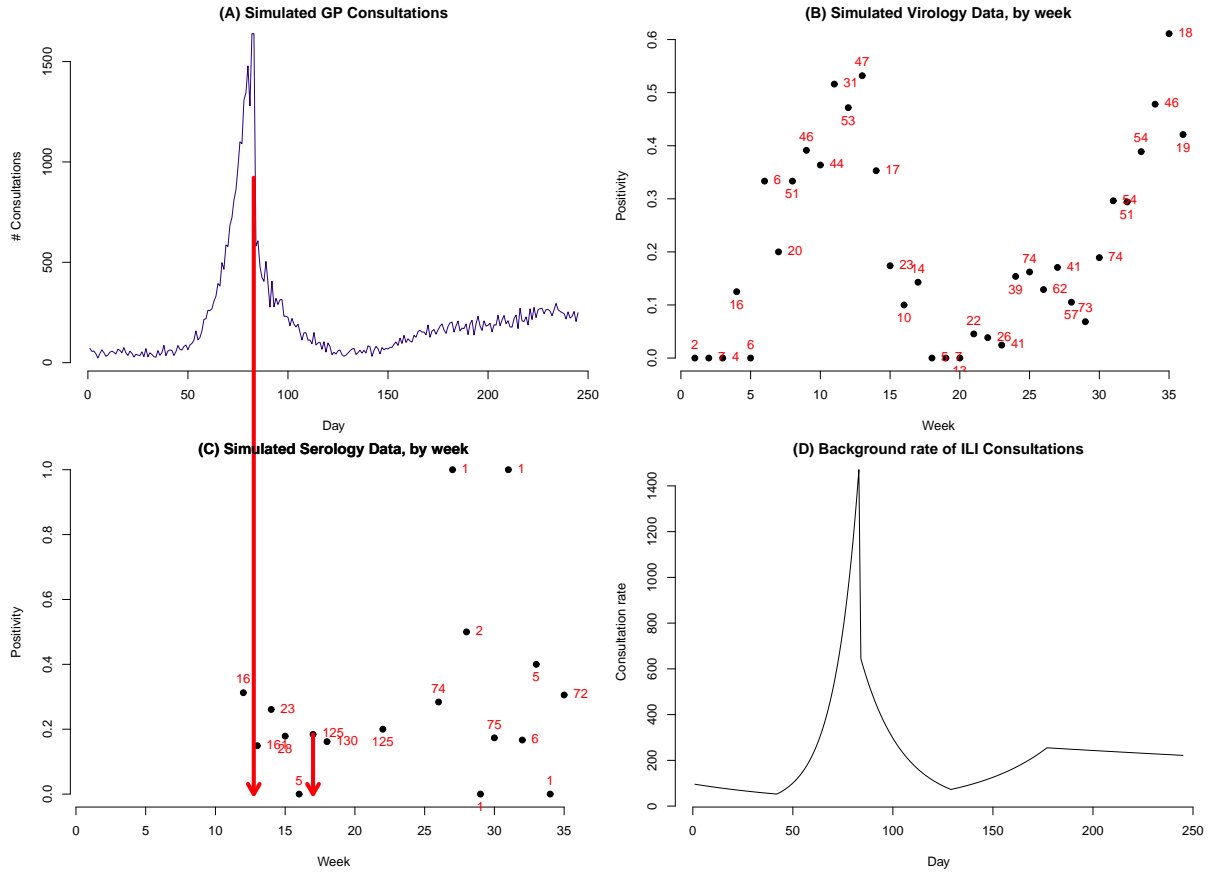


Figure 2: Top row: (A) Observed number of GP consultations $X_{t_k,a}^P$; (B) swab positivity data ($W_{t_k,a}$) with numbers representing the size of the weekly denominator. Bottom row: (C) serological data ($Z_{t_k,a}$); (D) pattern of background consultation rates for the primary healthcare data aggregated over ages. The red arrows over figures (A) and (C) highlight the timing of some key, informative observations

5 Results from a naive SMC algorithm

In this section we try and recreate the process of tracking the evolution of an epidemic, comparing the performance of a range of different SMC schemes to the pragmatic ‘gold-standard’ MCMC algorithm in Section 2.2.

The focus of a public health response in the early stages of an emerging epidemic will be on the estimation of some key epidemic parameters from a few initial cases in localised outbreaks. So, here, we assume that real time monitoring of the epidemic will only be required after this initial stage, which is taken here to be the first 50 days of the epidemic. For ease of presentation we introduce a number of staging posts over time, at $t_k = 50, 70, 83, 120, 164$ and 245 days, that reflect the changing characteristics of the epidemic. At each of these points, an MCMC implementation of the model will be carried out and the SMC algorithm will then be used to propagate the MCMC-obtained posteriors over the intervals defined by these staging posts. For example, the MCMC-obtained estimate $\pi_{50}^{\text{MCMC}}(\theta)$ of $\pi_{50}(\theta)$ will be used as the starting point for the SMC algorithm over the interval 50-70 days. The algorithm will be run for 20 days of data to give an estimate $\pi_{70|50}^{\text{SMC}}(\theta)$, which will then be compared with $\pi_{70}^{\text{MCMC}}(\theta)$. The similarity (or divergence) between the two distributions is measured by an approximation to the Kullback-Leibler (KL) divergence of $\pi_{t_k|t_k}^{\text{SMC}}(\theta)$ from $\pi_{t_k}^{\text{MCMC}}(\theta)$, obtained by assuming that both distributions are multivariate normal:

$$\begin{aligned} KL\left(\pi_{t_k}^{\text{MCMC}} \parallel \pi_{t_k|t_k}^{\text{SMC}}\right) &= \int_{\Theta} \pi_{t_k}^{\text{MCMC}}(\theta) \log \left\{ \frac{\pi_{t_k}^{\text{MCMC}}(\theta)}{\pi_{t_k|t_k}^{\text{SMC}}(\theta)} \right\} d\theta \\ &\approx \frac{1}{2} \left\{ \text{tr}(\Sigma_1^{-1} \Sigma_0) + (\mu_1 - \mu_0)^T \Sigma_1^{-1} (\mu_1 - \mu_0) - \dim(\theta) + \log \left(\frac{|\Sigma_1|}{|\Sigma_0|} \right) \right\} \end{aligned}$$

where $\pi_{t_k}^{\text{MCMC}}(\theta)$ and $\pi_{t_k|t_k}^{\text{SMC}}(\theta)$ are approximated by $N(\mu_0, \Sigma_0)$ and $N(\mu_1, \Sigma_1)$ multivariate normal distributions, respectively.

5.1 Scenario 1: Using Confirmed cases and Serology Data

Table 2 reports the KL divergences between the target posterior distributions obtained from SMC and MCMC at $t_k = 50, 70, 83, 120, 164$ and 245 days, for each of the SMC kernels discussed in Section 3.1. These results show that the correlated random-walk method (see (12)) is notably inferior over most of the intervals prior to the interval 84–120 days. Beyond this time, as data accumulate, the divergence between distributions π_k and π_{k+1} is small and the conservative random-walk proposals become progressively more adequate at bridging the gap between the two. The component-wise approximate Gibbs scheme (13) performs better (*i.e.* minimises KL divergence) over almost all intervals. As the move step here has many accept-reject steps (one for each component grouping of Table 1), each of low dimension, the overall acceptance rate (the proportion of particles for which at least one component moves) is close to one, giving $ESS \approx n_k$.

Figure 3 illustrates the performance of the approximate Gibbs component-wise proposal kernel comparing the SMC- and MCMC-obtained scatterplots for the parameter components ψ and v at $t_k = 70$ (A), $t_k = 120$ (B) and $t_k = 245$ (C). The grey points in both the left and the right panels represent the MCMC-obtained sample at the beginning of the interval, with the overlaid coloured points representing the SMC or MCMC-obtained samples at the end of the interval. In the SMC-obtained samples, the colour of the plotted points represents the weight attached to the particle, with the red particles being those of heaviest weight. It can be seen that there is close

Table 2: Kullback-Leibler statistics and likelihood evaluations per day ('Run Time') for each resample-move algorithm used to analyse the confirmed case and serology dataset over each of the time periods studied.

<i>Proposal Method</i>		<i>Correlated Random-Walk</i>	<i>Component-wise approx. Gibbs</i>	<i>Block approx. Gibbs</i>
<i>Intervals</i>				
0-50	KL	2.83	2.58	2.61
	Run Time	18200	16800	8000
51-70	KL	2.00	0.908	1.32
	Run Time	21000	21000	8000
71-83	KL	4.44	1.06	1.60
	Run Time	26923	26923	7692
84-120	KL	16.3	6.58	2.09
	Run Time	20811	17027	10000
121-164	KL	0.106	0.113	0.122
	Run Time	3182	3182	4773
165-245	KL	0.339	0.471	1.15
	Run Time	8642	9506	9136

Table 3: Kullback-Leibler statistics and likelihood evaluations per day (‘Run Time’) for each resample-move algorithm used to analyse the ‘contaminated’ primary care consultation and serology dataset over each of the time periods studied.

<i>Proposal Method</i>		<i>Correlated Random-Walk</i>	<i>Component-wise approx. Gibbs</i>	<i>Block approx. Gibbs</i>
<i>Intervals</i>				
51-70	KL	3.54	1.43	3.98
	Run Time	33000	27000	9500
71-83	KL	1.54	7.74	10.1
	Run Time	23100	18500	6920
84-120	KL	222	26400	17700
	Run Time	29200	30800	9460
121-164	KL	2.63	0.477	2.17
	Run Time	15000	15000	9320
165-245	KL	9.63	2.24	1.90
	Run Time	12600	12600	9750

correspondence between the SMC and MCMC obtained distributions at $t_k = 70$ and $t_k = 245$, but substantial departure at $t_k = 120$.

The results presented in Table 2 are based on just one SMC and MCMC simulation applied to a single set of simulated data. However, this process has been repeated numerous times, with results proving to be both qualitatively and quantitatively robust.

5.2 Scenario 2: Using Primary Care Consultations and Serology Data

With a parameter space that has expanded from 15 to 24 dimensions, the problems identified in Table 2 are magnified (see Table 3). In particular, after $t_k = 83$ the number of new parameters that become active is greater than in Scenario 1. For the following few days, some of these new parameters are only weakly identifiable. As a result, it can be seen from Table 3 that the KL divergences over the interval 83-120, irrespective of the proposal scheme, are arbitrarily high. The higher dimensionality of the particles making it even harder for the algorithm to find suitable areas of high posterior density with only a limited number of Metropolis-Hastings steps available.

5.3 Remarks

From the results presented above, it is clear that the naive SMC algorithm cannot handle the ‘shock’ in the count data occurring at $t_k = 83$. In scenario 1, this shock is accommodated by the model through step changes in the parameters, η_{t_k} , $p_{t_k,a}^e$ with similar step changes in the levels of background consultation in Scenario 2. For these parameter components, the target marginal posterior distributions move rapidly from day 84 as probability density shifts away from unin-

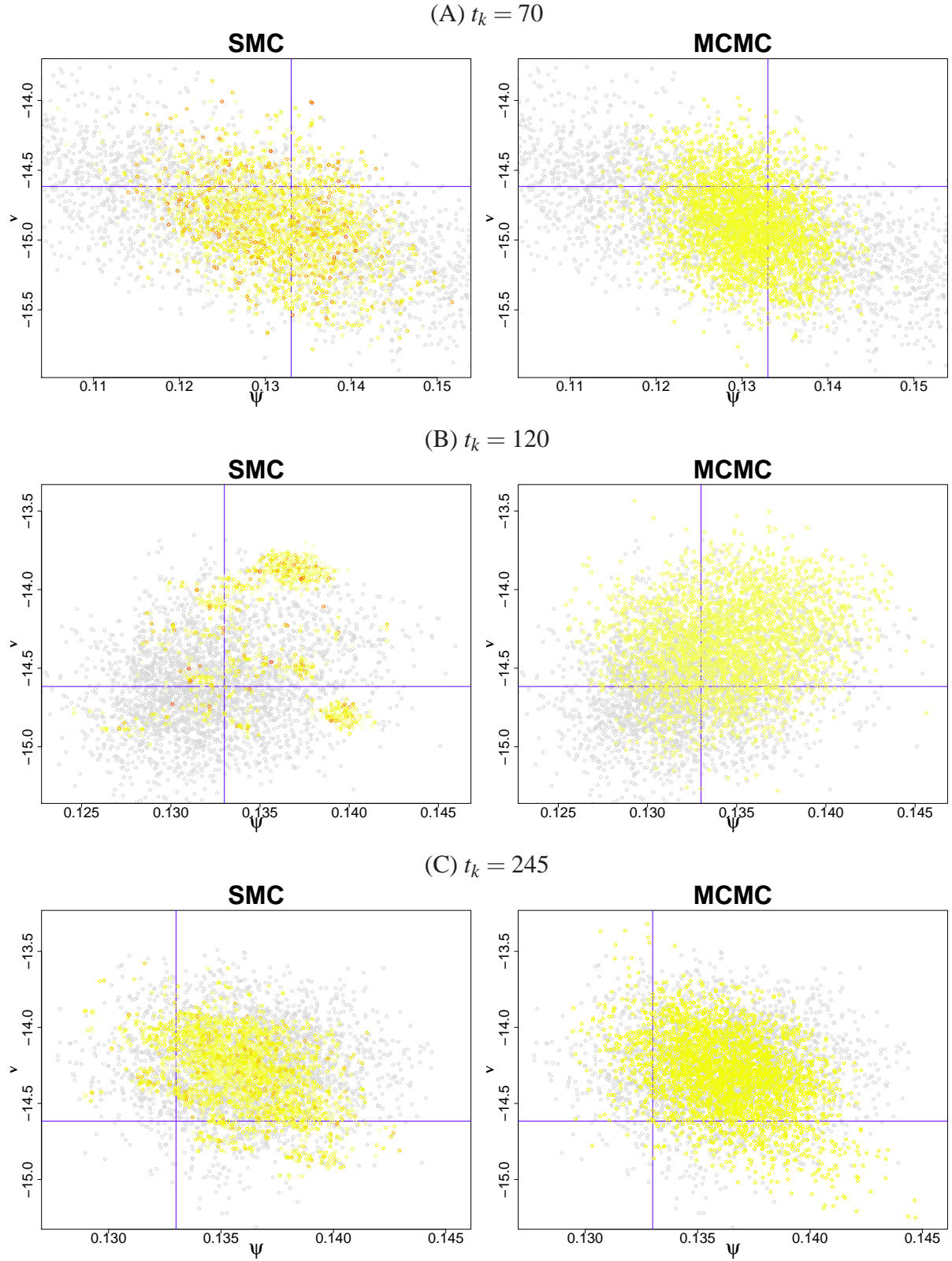


Figure 3: Comparison of SMC-obtained posteriors and MCMC-obtained posteriors at $t_k = 70$ (A), $t_k = 120$ (B) and $t_k = 245$ (C) days, via scatter plots for the parameters ψ and v

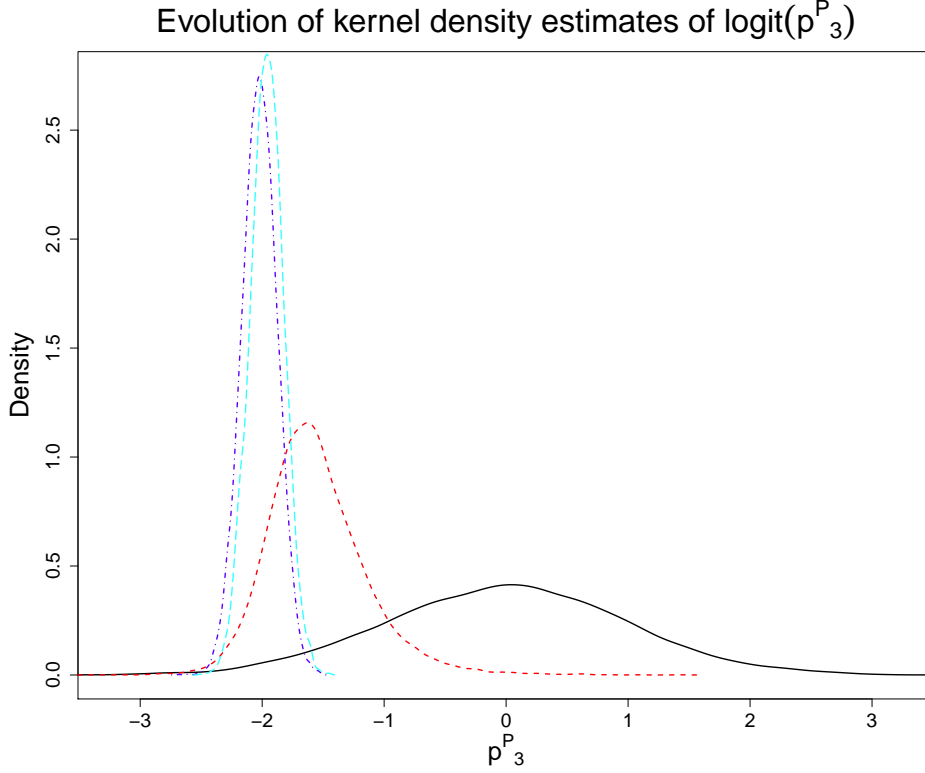


Figure 4: The movement of the marginal posterior for parameter p_3^P in the interval 84 – 120 days. The black line gives the density after 83 days, the dashed red, dark blue and light blue lines correspond to the posterior densities on days 84, 110 and 120.

formative prior distributions. As an example, Figure 4 shows the change in the marginal posterior for the parameter p^P over the interval [84, 120] days. This effect is particularly exaggerated for the overdispersion parameter η_{t_k} , for which the chosen $\Gamma(0.01, 0.01)$ prior distribution is unbounded and highly non-Gaussian even after transforming to the log-scale. In this case, there is very little prior to posterior overlap and normal proposal distributions can represent a poor choice.

For Scenario 1, the 84-120 day interval is the only one over which the block-update approximate Gibbs method gives the best performance (as measured by the KL divergence in Table 2). This is attributable to the way the proposal mechanisms modify the ESS as discussed in Section 3.1.1. The inability of the full-block Gibbs updates to restore the ESS, unlike the component-wise algorithms, leads to a rejuvenation step at the arrival of each new data point. This greater number of rejuvenations better enables this algorithm to track the shifting posterior distributions better over time, although it negates any advantages of this algorithm in terms of computation time (see Tables 2 and 3). However, even with the block updates, good correspondence between the SMC- and MCMC-obtained posteriors is not achieved in Scenario 1 until $t_k \approx 100$, and not at all in Scenario 2.

From these initial results it is clear that an alternative algorithmic formulation is needed for computationally efficient inference when target posteriors are highly non-Gaussian and/or are moving fast between successive batches of data as a consequence of highly informative observations.

6 Extending the Algorithm - Handling Informative Observations

A key feature of any improved SMC algorithm must be that the ESS (10) retains its interpretation as the “required size of an independent sample drawn directly from the target distribution to achieve the same estimating precision attained by the sample contained in the particle set” (Carpenter et al., 1999). A single step of the component-wise algorithms of Section 5 lead to the full or near-full restoration of the ESS, *i.e.* $ESS \approx n_k$. In general, the smaller the proposal scaling, the greater the number of accepted moves, the greater the number of unique particles and the higher the resulting value for the ESS. The limit of this is a highly clustered posterior sample, barely distinguishable from the set of resampled particles. Such a sample is not as informative as an independent sample of size n_k , and so the ESS, as calculated from the particle weights, is no longer a reliable guide to the quality of the sample. Similarly, the behaviour of the ESS (see Section 5.1) ceases to be an adequate guide for identifying rejuvenation times when considering block-update approximate-Gibbs proposals, as it is unable to recover to the threshold $\varepsilon_L n_k$.

We look at three possible improvements to the naive algorithm of Section 3, to ensure that the ESS remains a good measure of the quality of the sample: we address the timing of rejuvenations; we reconsider the choice of kernels used in the rejuvenations; and finally, we question the number of iterations we need to run the MCMC sampler before the sample is fully rejuvenated.

6.1 Timing the Rejuvenations: a Continuous-Time Formulation

Here we discuss the idea of rejuvenating at intervening times, including only a fraction of the new batch of data, to ensure that the capability to track the evolution of the posterior target distribution is retained.

In cases of large divergence between consecutive target distributions π_k and π_{k+1} , intermediate distributions can be constructed to allow the particle set to move gradually between the two targets (Del Moral et al., 2006). These intermediate distributions are generated via tempering (Neal, 1996), by gradually introducing the new batch of data into the likelihood at a range of temperatures $\delta \in [0, 1]$. These distributions are $\pi_{k,\delta}(\theta) \propto \pi_k(\theta) \{p(y_{k+1}|\theta)\}^\delta$.

Assume that batch of data y_{k+1} arrives uniformly over the $(k+1)^{th}$ interval rather than at the end of the interval, so that the particle weights develop according to

$$\tilde{\omega}_{k+\delta}^{(j)} = \omega_k^{(j)} \left\{ p(y_{k+1}|\theta^{(j)}) \right\}^\delta, \quad j = 1, \dots, n_k$$

More generally, denote $\omega_{k+\delta, \delta_0}^{(j)}$ the weight attached to a particle at an intermediate time $t_{k+\delta}$ when the previous rejuvenation took place at time $t_{k+\delta_0}$, with $\delta_0 = 0$ corresponding to no prior rejuvenation within the interval $(t_k, t_{k+1}]$. Then, for $\delta \geq \delta_0$,

$$\tilde{\omega}_{k+\delta, \delta_0}^{(j)} \propto \begin{cases} \left\{ p(y_{k+1}|\theta^{(j)}) \right\}^{\delta-\delta_0} & \delta_0 > 0 \\ \omega_k^{(j)} \left\{ p(y_{k+1}|\theta^{(j)}) \right\}^\delta & \delta_0 = 0 \end{cases}.$$

Therefore, if $ESS(\{\tilde{\omega}_{k+1, \delta_0}^{(j)}\}_{j=1}^{n_k}) < \varepsilon_L n_k$ a further rejuvenation would be proposed at time δ^* , such that $\delta^* = \arg \min_{\delta \in (\delta_0, 1)} \{ESS(\tilde{\omega}_{k+\delta, \delta_0}^{(j)}) - \varepsilon_L n_k\}^2$. This introduces the potential of making

a number of MH steps within the $(k + 1)^{\text{th}}$ interval that is in proportion to the degree of particle impoverishment that would have occurred had no rejuvenation steps been taken.

6.2 Choosing Kernels - Hybrid Algorithms.

As discussed in Section 5.3 each of the possible MH kernels has its own distinct strengths. In the construction of the MH-steps, these strengths can be exploited by using a combination of kernels. Full block approximate-Gibbs updates are efficient at reducing the clustering that forms around resampled particles. Adding a random walk step would allow the proposal of values outside the space spanned by the principal components of $\bar{\Sigma}_k$, something of particular necessity if the ESS becomes very small and $\bar{\Sigma}_k$ is close to singularity.

This motivates a hybridisation of the proposal mechanism. This can be done either by using mixture proposals, *e.g.* a mixture between the approximate Gibbs' proposals and full block ordinary random walk Metropolis proposals (Kantas et al., 2014), or, as in this case, by augmenting full block approximate Gibbs updates with componentwise random walk proposals. These will be used throughout the remainder of this section.

6.3 How Many MH Iterations? Multiple Proposals and Intra-class Correlation

In the MH-step of the algorithm, we are effectively working with n_k parallel MCMC chains. To keep making proposals until all the chains have attained convergence would be an inefficiency. In theory, the distribution governing the starting states of these MCMC chains forms a biased sample from the stationary distribution of the MCMC chain. It then seems a reasonable and sufficient requirement that we carry out MH steps until the chains have, to some degree, collectively 'forgotten' their starting positions. This can be monitored through an estimate of an intra-class correlation coefficient, ρ . To get such an estimate, we divide the particle set into I clusters, each of size $d_i, i = 1, \dots, I$, according to the parent particle at the resampling stage. For example, if a particular particle is resampled 5 times, it defines a cluster in the new sample with $d_i = 5$. Clusters with $d_i \leq 1$ are omitted as they have no within cluster variation (Donner and Koval, 1980). Furthermore, we pick univariate summaries $g_{ij} = g(\theta_{ij})$ of the epidemic curve described by the j^{th} particle in the i^{th} cluster, θ_{ij} . The 'attack rate' of the epidemic, the cumulative number of infections caused by the epidemic, is an obvious choice as it is a key measure of epidemic burden. Formally, the attack rate is defined:

$$g(\theta) = \frac{\sum_{t=1}^{\infty} \sum_{a=1}^A \Delta_{t,a}(\theta)}{\sum_a N_a}. \quad (14)$$

With classes of size $d_i, i = 2, \dots, I$, ρ can be estimated by the analysis of variance intra-class correlation coefficient (Donner and Koval, 1980; Sokal and Rohlf, 1981), given by

$$r_A = \frac{(MS_a - MS_w)/d_0}{(MS_a - MS_w)/d_0 + MS_w},$$

where $MS_a = \frac{1}{I-1} \sum_{i=1}^I d_i (\bar{g}_{i\cdot} - \bar{g}_{\cdot\cdot})^2$, $MS_w = \frac{1}{d-1} \sum_{i=1}^I \sum_{j=1}^{d_i} (g_{ij} - \bar{g}_{i\cdot})^2$ and $d_0 = \bar{d} - \frac{1}{d(I-1)} \sum_{i=1}^I (d_i - \bar{d})^2$, represent a between-class mean sum-of-squares, a within-class mean sum-of-squares and an average class size respectively, with $d = \sum_{i=1}^I d_i$, $\bar{g}_{i\cdot} = \frac{1}{d_i} \sum_{j=1}^{d_i} g_{ij}$ and $\bar{g}_{\cdot\cdot} = \frac{1}{d} \sum_{i=1}^I \sum_{j=1}^{d_i} g_{ij}$. Prior to the MH-phase of the algorithm, r_A will be equal to 1, as there is no within-class variation. However, with each iteration of the chosen MH-sampler, ρ will decrease and, in general,

so will its estimate r_A . We aim to choose a sufficiently small positive threshold for r_A to be the point beyond which there is no longer any value in carrying out further MH proposals to rejuvenate the sample, as particles spawned from different progenitors become indistinguishable from each other. Ideally we will choose this threshold to be as large as is practicably possible to minimise the number of rejuvenations required. We shall test our algorithms with thresholds $r_A^* = 0.1, 0.2, 0.5$.

7 Results from an Adapted SMC Algorithm

In Sections 5.1-5.2, results were presented for a number of time intervals. Now we focus solely on the 83 – 120 days interval, which contains the highly informative data discussed. In what follows, a hybrid algorithm is adopted, using combinations of three thresholds for r_A with both the continuous and discrete sequential algorithms.

7.1 Scenario 1: Using Confirmed Case and Serology Data

7.1.1 Choosing an Algorithm: Kullback-Leibler

Continuing to analyse the simulated data (Section 5.1), further MCMC samples were obtained using data up to and including days 84, 85, 86, 87, 90, 100, 110 and 120, so that KL divergences could be computed at each of these additional time points. In Figure 5(A), KL discrepancies are plotted over time for each combination of algorithms and thresholds.

From Figure 5(A) the lower value of r_A^* , the closer the SMC approximation to the MCMC-obtained posterior distributions. This is to be expected as a low value of r_A^* requires a greater number of iterations of the MCMC chain within each rejuvenation (see Figure 5(C)), leading to SMC-derived posteriors more closely resembling the ‘gold-standard’.

It is difficult to interpret the KL divergences in their own right. To construct a reference distribution of such divergences, the MCMC analyses were repeated a further 40 times at each of the times in Table 4, using different starting points and random seeds. The KL divergences of each of these 40 posterior distributions from the original MCMC analysis were then calculated. We look for divergences in Table 4 that are typical of this sample of KL divergences. The cells in the table highlighted in green indicate where the tabulated KL divergence lies among the lower 95% of sampled KL values at that time, a threshold marked in Table 4 as the ‘KL target’.

From Table 4, for $t_k = 84, 85, 86$, there is no clear best performing algorithm as none manage to yield a value for the KL that is typical of an MCMC analysis. From $t_k = 87$ onwards, the continuous-time algorithm is much more efficient as for the given r_A^* , fewer iterations are required to attain a much lower KL value (see Figures 5(A) and 5(C)). These results were tested under multiple reruns of the algorithm. For $r_A^* = 0.5$, findings from the continuous-time algorithm were highly volatile. Despite this, Figure 5(B) shows that the number and timing of rejuvenations appears independent of the choice of r_A^* and the decline in the ESS is independent of the quality of the initial sample. In conclusion, the continuous-time algorithm is to be preferred with a threshold of, at most, 0.2. Note that for $r_A^* = 0.1$, the continuous-time algorithm has KL typical of an MCMC analysis for all values of $t_k \geq 87$.

7.1.2 Acceptance Rates

Performance of the continuous-time algorithm appears strongly linked to the acceptance rate of the block approximate Gibbs’ proposals. This acceptance rate is particularly low prior to

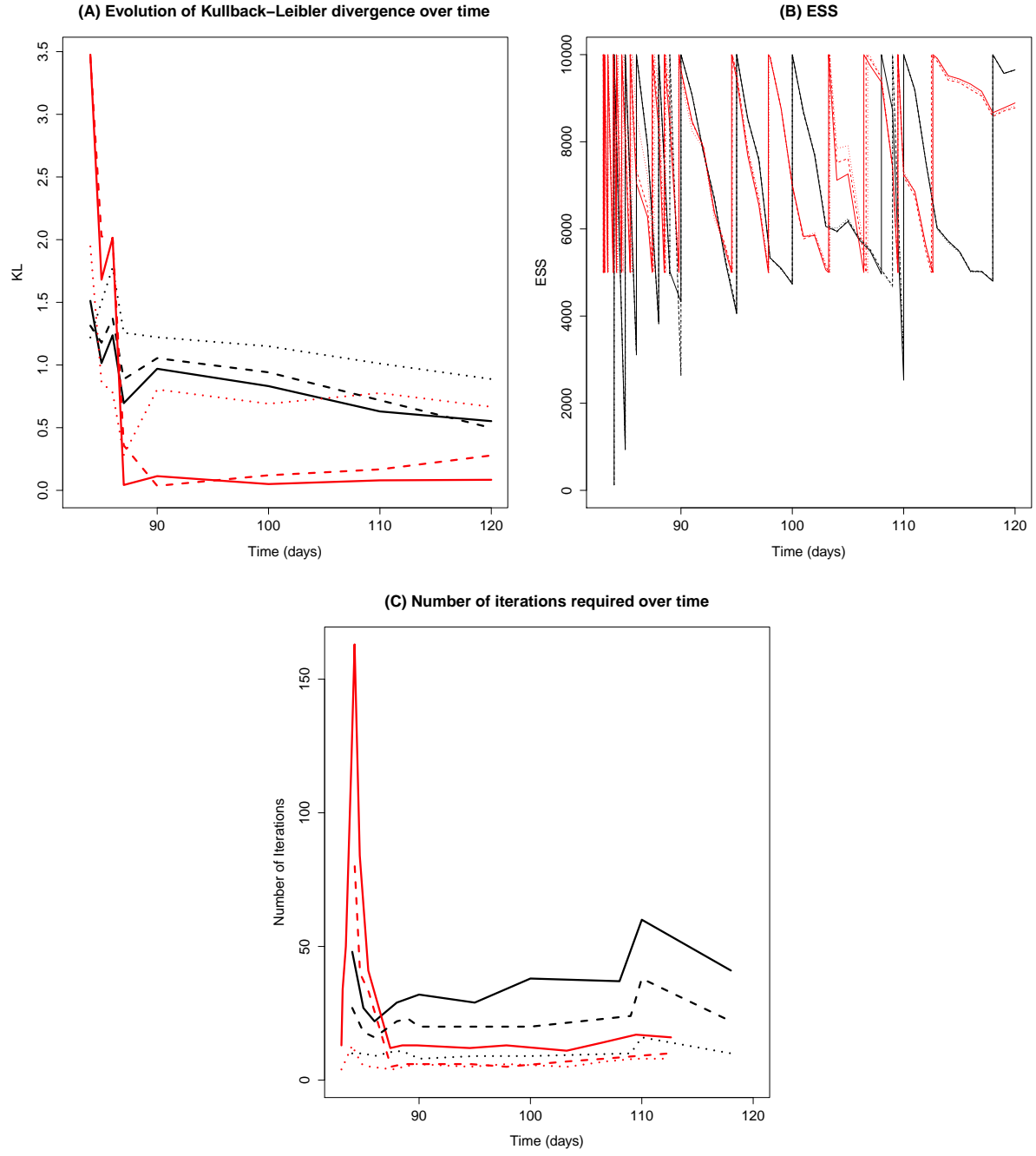


Figure 5: (A) Kullback-Leibler divergence over time; (B) ESS for different thresholds using both discrete and continuous time algorithms; (C) Number of proposals required at each rejuvenation time by algorithm. Black and red lines correspond to the use of discrete and continuous algorithms respectively, solid, dashed and dotted lines correspond to the use of the thresholds $r_A^* = 0.1, 0.2$ and 0.5 respectively.

Table 4: Performance of the adapted SMC algorithm over the interval 83-120 days by ICC threshold. Traditional filter using daily data ('discrete'); continuous-time filter ('continuous'). Cells shaded in green contain KL divergence that is lower than the reference KL divergence distribution.

<i>ICC threshold</i>	<i>0.5</i>	<i>0.2</i>	<i>0.1</i>	<i>ICC threshold</i>	<i>0.5</i>	<i>0.2</i>	<i>0.1</i>
84 Days (KL target = 0.732)				90 Days (KL target = 0.159)			
Continuous	1.95	3.46	3.48	Continuous	0.805	0.0358	0.113
Discrete	1.22	1.31	1.51	Discrete	1.22	1.05	0.970
85 Days (KL target = 0.135)				100 Days (KL target = 0.135)			
Continuous	0.862	2.03	1.68	Continuous	0.691	0.120	0.0501
Discrete	1.50	1.18	1.02	Discrete	1.15	0.942	0.832
86 Days (KL target = 0.365)				110 Days (KL target = 0.122)			
Continuous	0.780	2.01	2.02	Continuous	0.776	0.167	0.0799
Discrete	1.78	1.37	1.24	Discrete	1.01	0.719	0.630
87 Days (KL target = 0.276)				120 Days (KL target 0.119)			
Continuous	0.282	0.358	0.0427	Continuous	0.666	0.278	0.0842
Discrete	1.26	0.887	0.696	Discrete	0.888	0.498	0.552

$t_k = 87$, before undergoing a sudden step change and increasing from 1-2% to 15-20%, although, for $r_A^* = 0.5$, this step change only occurred in about 50% of runs. In contrast, the acceptance rates for the discrete-time algorithm are consistently around 5% throughout, as seen from the number of proposals required over time (Figure 5(C)). The result of this is that, from day 87 onwards, far fewer proposals are required in total for the continuous-time algorithm, even if the number of rejuvenation times increases.

7.1.3 Parameter Estimation

To visualise the performance of the SMC algorithms, Figures 6 and 7 contrast SMC- and MCMC-obtained joint posteriors using scatterplots similar to those in Figure 3. Plots refer to the joint distribution of three pairs of parameters obtained from the discrete-time (Figure 6) and continuous-time algorithm using $r_A^* = 0.2$ at $t_k = 84, 87$ and 120 days (Figure 7). The top pair of parameters corresponds to two well-estimated (prior to $t_k = 83$) parameters, ψ and v , the epidemic growth rate and the log-initial number of infectives. The middle pair correspond to the dispersion parameter beyond $t_k = 83$, η_2 , and m_4 , the reduction in the contacts amongst 5-14 year-olds due to the over-summer school holiday. These parameters are new, or nearly new, additions to the likelihood around $t_k = 83$, and significant departure from their prior distributions is expected over the interval 84-120 days. Finally, the bottom pair, (p_3^C, p_4^C) , are also new to the likelihood, and therefore also departing from their prior distributions after $t_k = 83$. They correspond to the propensity for both adults and children to be confirmed in the new reporting regime after $t_k = 83$ (see Table 1). Both figures show that SMC and MCMC posteriors are quite similar, with some possible irregularity in Figure 6 in the contours for the supposedly well-estimated parameters ψ and v , indicative of the presence of clustering that is not present in the MCMC analysis at times $t_k = 84, 87$. This clustering is distinctly less evident in the plots corresponding to the continuous-time algorithm.

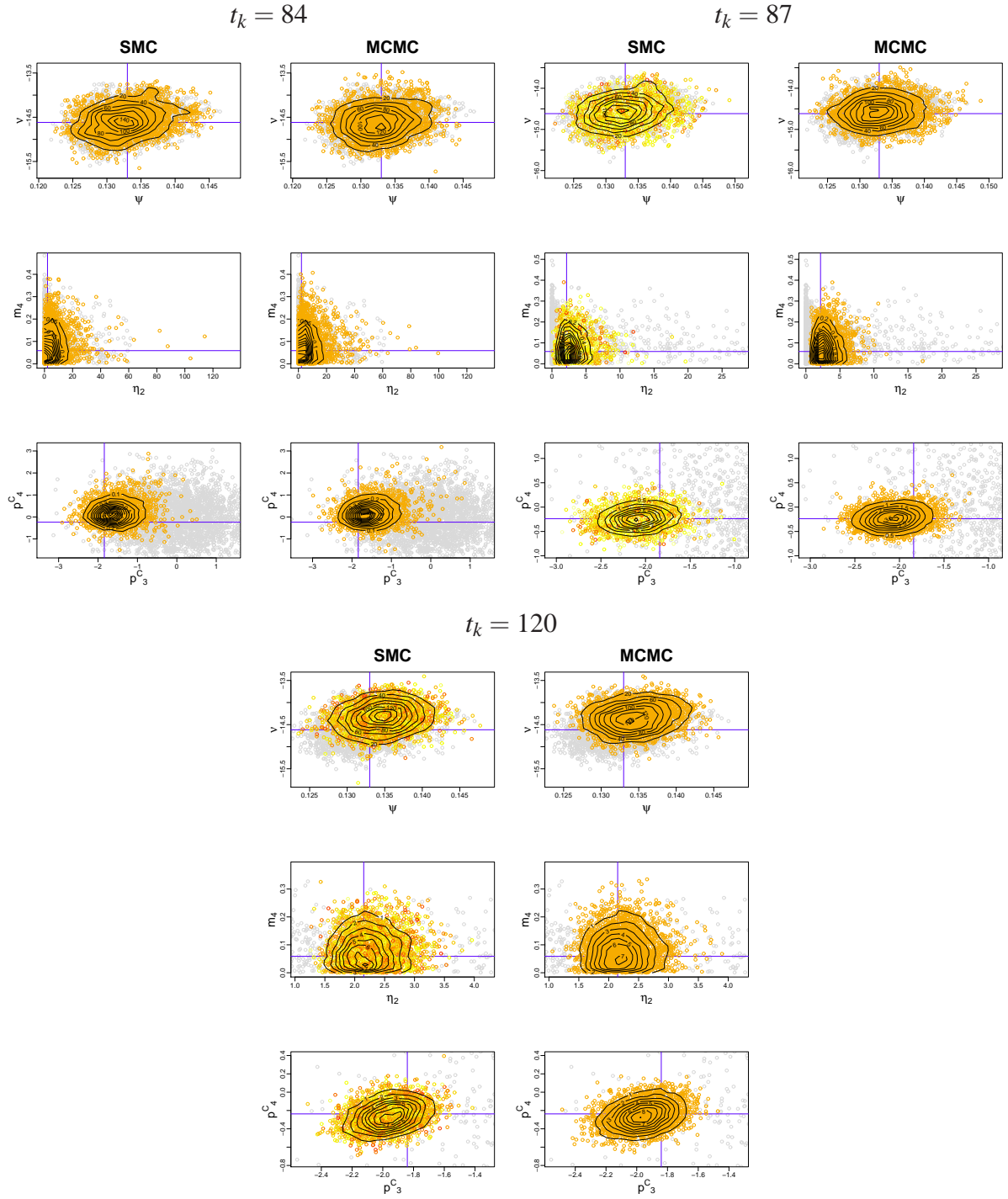


Figure 6: Posterior scatters and contours for three pairs of parameters at times $t_k = 84$, $t_k = 87$ and $t_k = 120$ for the discrete-time algorithm with $r_A^* = 0.2$

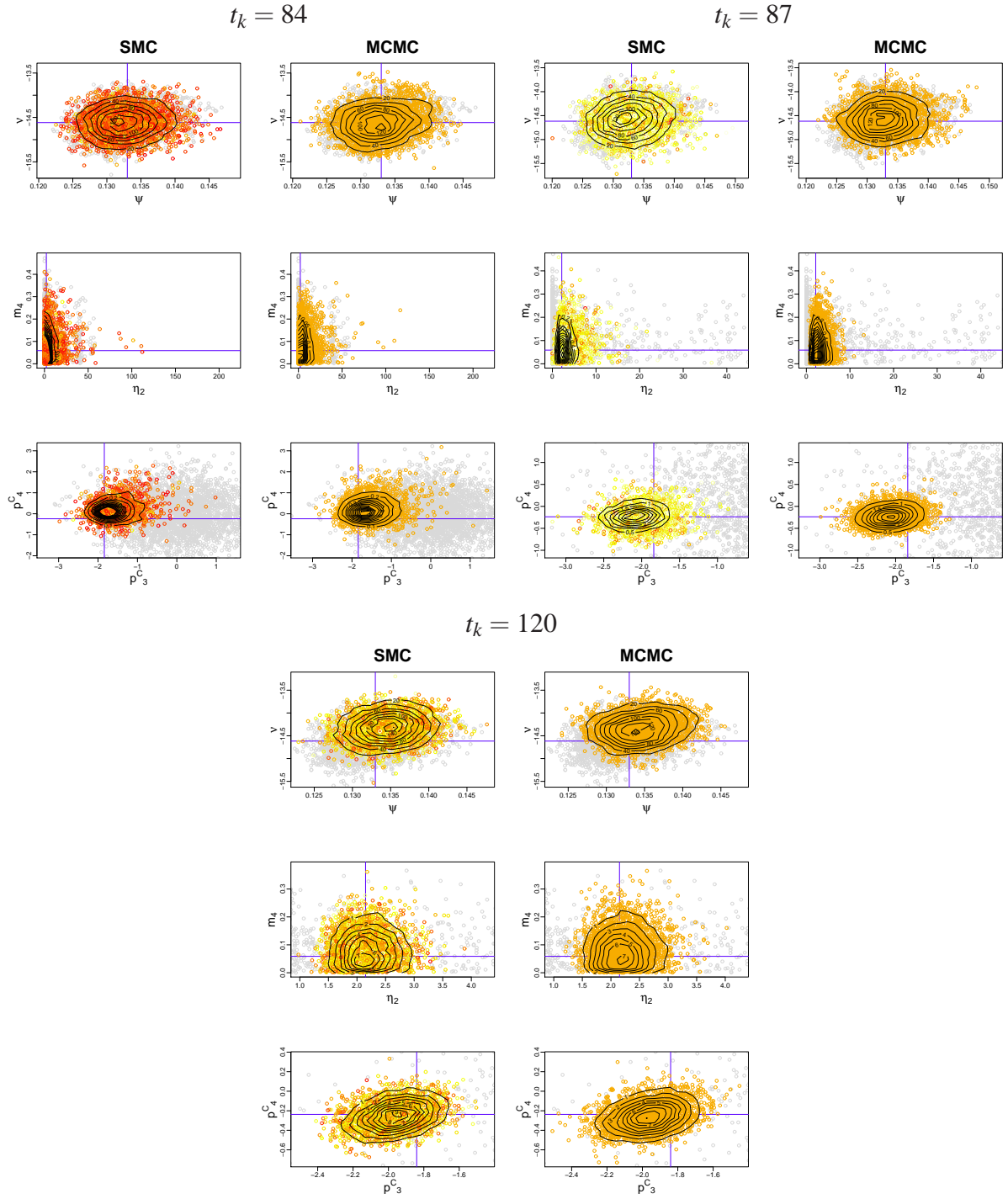


Figure 7: Posterior scatters and contours for three pairs of parameters at times $t_k = 84$, $t_k = 87$ and $t_k = 120$ for the continuous-time algorithm with $r_A^* = 0.2$

Table 5: Performance of the adapted SMC algorithm over the interval 83-120 days, using the continuous filter and the continuous filter alternative with the negative binomial dispersion parameters removed from the block proposals. Here, the parameters describing the background rates of consultation have been removed from the KL calculations.

<i>ICC threshold</i>	<i>0.5</i>	<i>0.2</i>	<i>0.1</i>	<i>ICC threshold</i>	<i>0.5</i>	<i>0.2</i>	<i>0.1</i>
84 Days (KL target = 6.06)				90 Days (KL target = 0.120)			
Continuous	2.92	2.87	2.83	Continuous	1.80	0.353	0.0663
Cts. Reduced	2.97	2.85	2.86	Cts. Reduced	2.10	0.0927	1.42
85 Days (KL target = 1.90)				100 Days (KL target = 0.182)			
Continuous	3.05	3.00	2.98	Continuous	0.157	0.102	0.0890
Cts. Reduced	3.06	2.97	2.98	Cts. Reduced	0.107	0.0835	0.0701
86 Days (KL target = 1.94)				110 Days (KL target = 0.0936)			
Continuous	3.28	3.24	3.25	Continuous	0.159	0.0774	0.111
Cts. Reduced	3.27	3.22	3.26	Cts. Reduced	0.197	0.0373	0.0348
87 Days (KL target = 5.44)				120 Days (KL target = 0.101)			
Continuous	2.54	2.45	2.42	Continuous	0.136	0.0435	0.0708
Cts. Reduced	2.51	2.48	2.44	Cts. Reduced	0.0999	0.0423	0.0551

7.2 Scenario 2: Using Primary Care Consultations and Serology Data

7.2.1 Choosing an Algorithm

As in Section 5.2, results from SMC applied to contaminated count data display many of the phenomena already discussed. Table 5 gives results comparable to those in Table 4 with cells highlighted in green to be interpreted as above.

The first thing to note in Table 5 is the absence of results from the discrete-time algorithm. On observing data at $t_k = 84$, the ESS falls from 10^4 to 2.24. From such an impoverished sample, the estimates $\bar{\theta}_{84}$ and $\bar{\Sigma}_{84}$, on which the proposal distributions in (12) and (13) depend, are highly variable. The resulting proposal distributions are unlikely to replicate the characteristics of the target posterior, $\pi_{84}(\cdot)$, and $\bar{\Sigma}_{84}$ can give conditional covariance matrices that are computationally singular. These factors conspire to give SMC algorithms that do not, within a reasonable amount of time, attain r_A^* . Therefore, the discrete-time algorithm has been dropped from consideration.

The second thing to notice is the addition of results from an algorithm labelled ‘cts.reduced’. Figures 8(A) and 8(B) display the kernel density estimates for the posterior marginal distributions of dispersion parameter η_2 and $\log(\eta_2)$ (over days 84-87 and again at day 90) showing that the distribution of η_2 is highly non-Gaussian after $t_k = 83$ days. It is only at day 90 that we see a distribution that resembles the normal distribution. This non-normality leads to very poor acceptance rates for the approximate-Gibbs’ proposals over the interval 83-89 days and stems from the uninformative gamma priors placed upon the η parameters. To improve acceptance rates the ‘cts.reduced’ algorithm was devised. The negative binomial dispersion parameters (see below) are omitted from the block approximate-Gibbs updates and are proposed separately. In terms of the resulting KL divergences, there is no significant drop in performance in moving from the continuous to the ‘cts. reduced’ algorithm as seen by the amount and position of the green cells in Table 5. The ‘cts. reduced’ proposal scheme, however, requires far fewer iterations of the Metropolis-Hastings algorithm over the interval 84-90 days. Figure 8(C) and (D) show the number of MH iterations required per rejuvenation and per day for both the ‘continuous’ vs. the ‘cts. reduced’ algorithms.

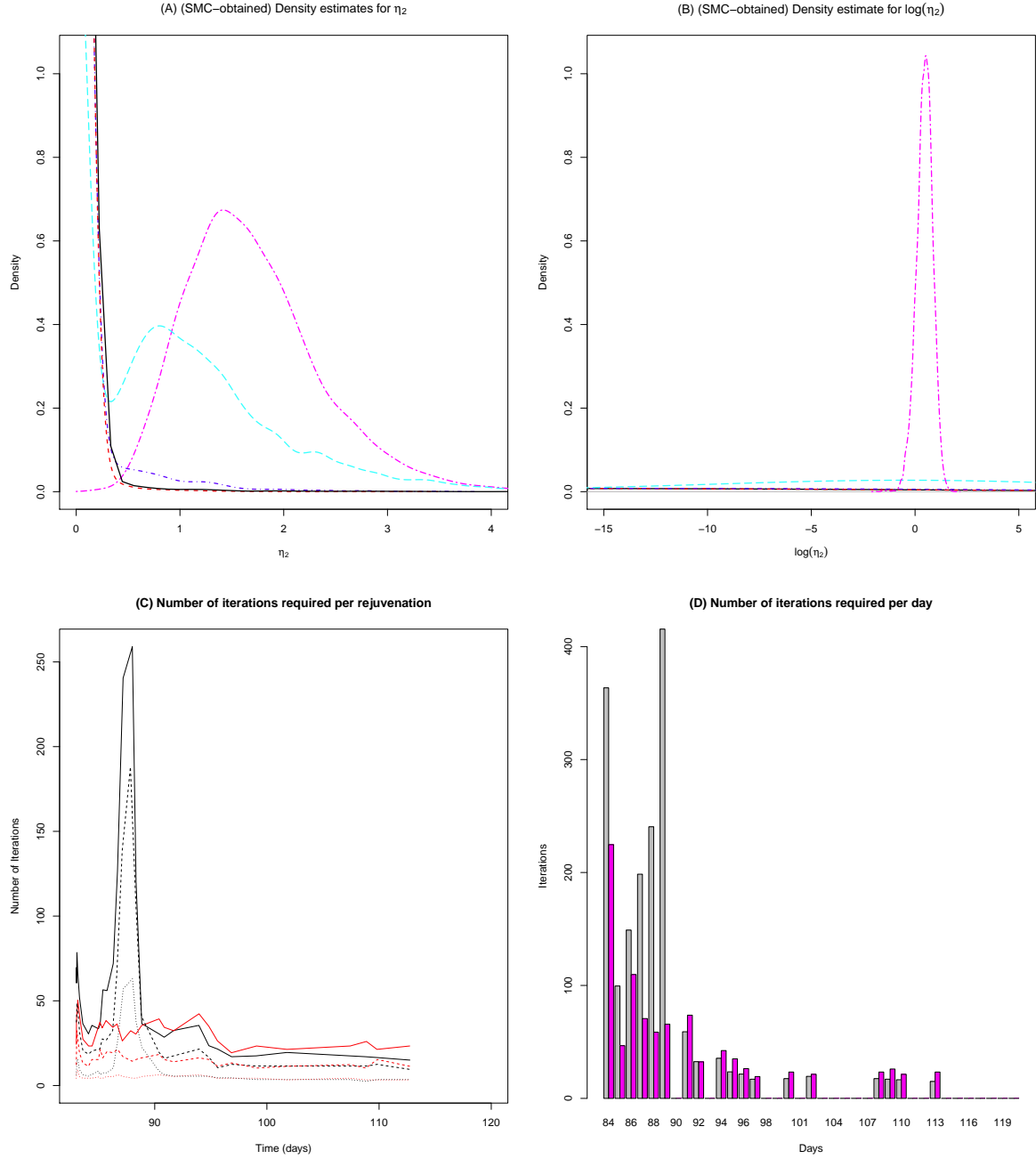


Figure 8: (A) and (B) Evolution of the densities of η_2 and $\log(\eta_2)$ after $t_k = 83$. The densities for the days 84, 85, 86, 87 and 90 days are solid black, and dashed red, dark blue, light blue and magenta; (C) Number of MH-steps required by the continuous-time SMC algorithms per rejuvenation against the timing of the rejuvenation for both the continuous time algorithms (black and red correspond to with and without η in the updates) for values of $r_A^* = 0.1$ (solid line), 0.2 (dashed line) and 0.5 (dotted line); (D) Total number of MH-steps required by the continuous-time SMC algorithms per time interval, with $r_A^* = 0.1$ and using the continuous-time algorithm (grey bars) and the same algorithm without η (magenta bars).

7.2.2 Acceptance rates

The number of iterations required by the ‘cts. reduced’ algorithm, reasonably consistent over the timepoints, correspond to an acceptance rate of about 10% for the full-block (excluding the η parameters) updates. For the continuous-time algorithm, these acceptance rates, though initially adequate, drop as low as 0.3% on day 89. This is shown by the peak of over 250 proposals per rejuvenation and over 400 proposals per day in Figures 8(C) and (D) respectively. Note that over time, as the target distribution converges to a multivariate normal distribution, the number of moves required for both methods equalise (in fact, the plain continuous-time algorithm is marginally faster) and there is no longer a benefit in using the ‘cts. reduced’ proposal scheme.

7.2.3 Parameter Estimation and Epidemic Projection

Most of the scatter plots contrasting the posterior distributions obtained under either the ‘continuous’ or ‘cts. reduced’ schemes show a similar level of correspondence to their MCMC-obtained counterparts to that observed in Figure 7 for the ‘continuous’ algorithm. However, for the parameters of the background consultation rates (see Figure 2(D)) this is not the case at all timepoints. For the first couple of days post $t_k = 83$ days, some of the parameters describing $B_{t_k,a}$ are only weakly identifiable. Figure 9 highlights this with scatterplots for two weakly identifiable parameter components β_3^B and β_9^B showing a clear discrepancy between the MCMC- and the SMC-obtained posterior scatters. The SMC distributions, being based on many short MCMC chains, cover the full posterior distribution adequately. For $t_k = 85, 86$, however, the MCMC has difficulty mixing, and this manifests in a particle scatter that is stuck in a sub-region of the full marginal support. KL discrepancies calculated for days where this weak identifiability exists (and it diminishes over time), will therefore be unreliable.

8 Discussion

This paper addresses the substantive real world problem of online tracking of an emergent epidemic, assimilating multiple sources of information through the development of a suitable SMC algorithm. When incoming data are stable, this process can be automated using standard SMC algorithms, confirming current knowledge (*e.g.* Dukic et al., 2012; Ong et al., 2010). However, in the likely presence of interventions or any other event that may provide a system shock, it is necessary to adapt the algorithm appropriately. On observing the impact that a new batch of data has on the ESS of a particle set, tailoring of the MH-kernel and selection of suitable thresholds can ensure efficient performance. However, as we have seen, given that not all prior distributions are well chosen and not all models well conceived this might necessitate some careful, yet ad hoc tinkering. The end result is an algorithm that is a hybrid of particle filter and population MCMC (Geyer, 1991; Liang and Wong, 2001; Jasra et al., 2007).

Having simulated an epidemic where a public health intervention provides a sudden change to the pattern of case reporting, we have constructed a more robust SMC algorithm by tailoring

1. the choice of rejuvenation times through tempering;
2. the choice of the MH-kernel by hybridising local random walk and Gibbs proposals; and
3. by introducing the use of the intra-class correlation to provide a stopping rule for the MCMC steps to limit the number of MCMC steps within each rejuvenation.

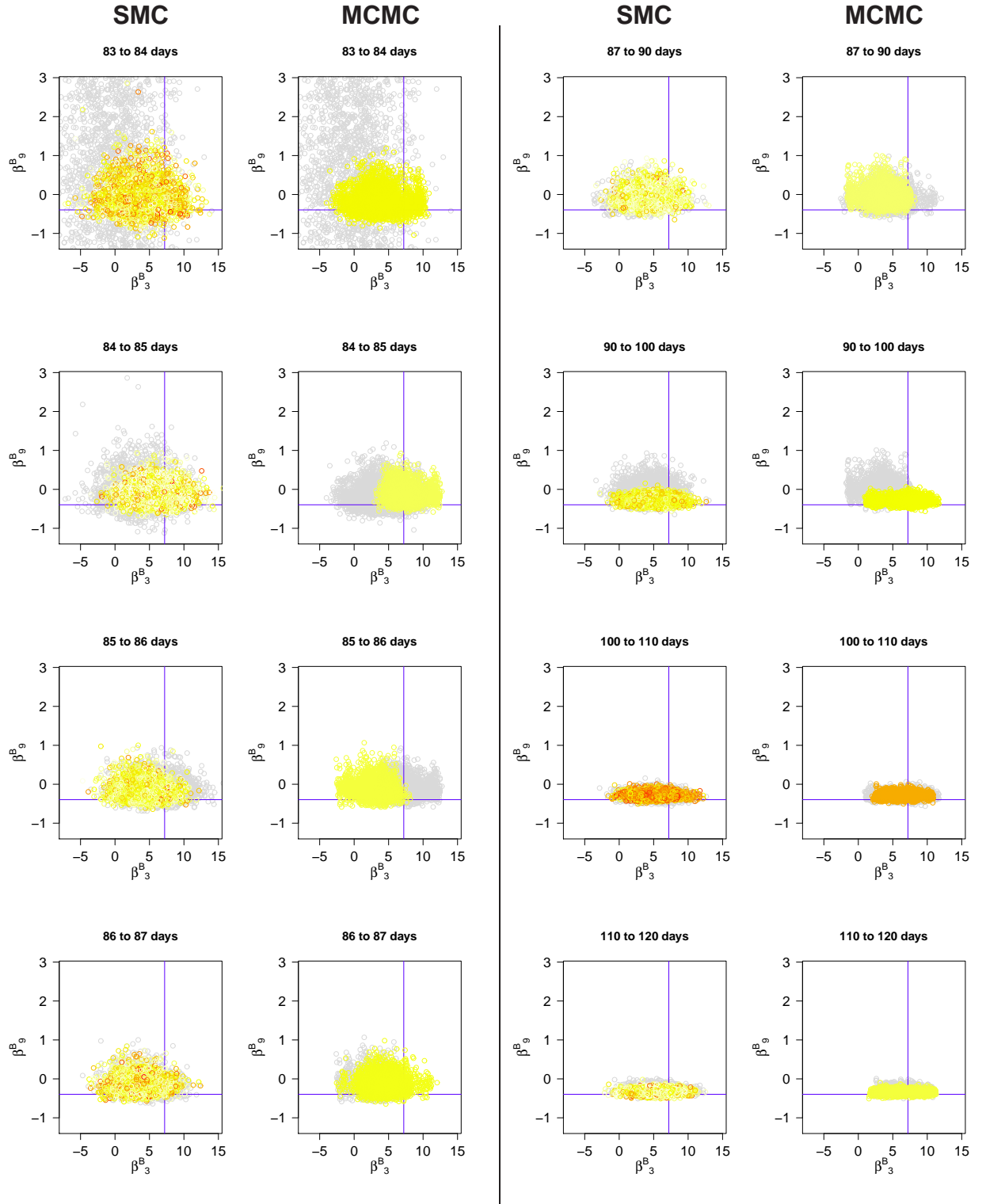


Figure 9: The evolution over time of the marginal joint posterior for two components of the parameter vector β^B . Comparison between SMC-obtained and MCMC-obtained posterior distributions. Grey points indicate the distribution at the start of the interval.

Our experience suggests, real time epidemic tracking will involve switching between a simple, automated, SMC to an SMC specifically tailored to the nature of any impending shock. Throughout we have inevitably made pragmatic choices and alternative strategies could have been adopted. We reflect on these, lessons learned and outstanding questions in what follows.

8.1 Rejuvenation at times of shocks

In the motivating example, a system “shock” occurred at $t_k = 83$. This shock represents a systematic change in the way the data are generated, affecting a number of parameters that, at this time, have a step-change in their values. The first few observations in the new parametric regime after the shock typically cause the greatest disturbance to the marginal posterior distribution for these parameters. Posterior $\pi_k(\theta)$ is no longer a good importance distribution to sample from $\pi_{k+1}(\theta)$ and proposal kernels based on a reweighted sample from $\pi_k(\theta)$ may not be useful. This will be reflected in a severe drop in the ESS.

A low value for the ESS is always indicative of depletion, whereas a high value does not guarantee that the sample is adequate. Section 5 illustrated how the ESS can be artificially rejuvenated even when the particle set is not. For the ESS to be useful, it is essential that previous rejuvenation steps result in a sufficiently independent set of values for the margins of interest. After resampling, many MH-steps may be needed to remove any clustering. This motivates the use of the analysis of variance intra-class correlation coefficient, r_A , to define a stopping rule for the MH-steps. Currently this rule relies on two algorithmic choices: the choice of a univariate function of interest, $g(\cdot)$ (see Equation (14)), and the choice of the threshold r_A^* , the largest acceptable value for r_A at the end of the rejuvenation process.

The function $g(\cdot)$ should depend on model outputs of particular relevance. The predicted attack rate of an epidemic is a quantity that will be reported to public health policymakers throughout an epidemic and is dependent on all the transmission parameters. However, when the parameter vector is high-dimensional, as in this case, is it reasonable to condense this into a univariate summary to use as a basis for a stopping rule? Convergence of MCMC is typically diagnosed by looking at marginal distributions, so should we be doing something similar here? Does this necessitate the use of multivariate analogs for the intra-class correlation coefficient (for example, see Ahrens, 1976; Konishi et al., 1991)? It is felt here that the univariate g is adequate as the parameters introduced at the ‘shock’ time are largely nuisance parameters not strongly correlated with the transmission parameters that influence g .

Once r_A has been suitably defined, a suitable stopping threshold, r_A^* has to be chosen. Given the antecedent prescription for defining clusters used here, then r_A truly is a measure of how well the particles have collectively ‘forgotten’ their starting points. In situations where the target posterior is well-matched by its Gaussian approximation, we could use a higher threshold than when starting from a poor estimate for the target distribution. A value of $r_A^* = 0.1$ is a sufficiently small threshold except for extreme cases of departure between two successive distributions.

A possible alternative to r_A is an extension of the sampling variance $\hat{V}_k(g)$ in Gilks and Berzuini (2001)

$$\hat{V}_k(g) = \frac{1}{\left(\sum_{m=1}^{n_k} \omega_k^{(m)}\right)^2} \sum_{m_1=1}^{n_k} \sum_{m_2=1}^{n_k} C_k^{m_1, m_2} \omega_k^{(m_1)} \omega_k^{(m_2)} \left(g_k^{(m_1)} - \bar{g}_k^*\right) \left(g_k^{(m_2)} - \bar{g}_k^*\right).$$

where $C_k^{m_1, m_2}$ gives the number of common ancestors of particles m_1 and m_2 within the interval, and $g_k^{(m)} = g(\theta_k^{(m)})$ is the function of interest, with \bar{g}_k^* an estimate of $\mathbb{E}_{\pi_k}\{g(\theta_k)\}$. $\hat{V}_k(g)$ was initially proposed to identify suitable rejuvenation times, but it is not clear how this can be done

prospectively. It could, however, provide a stopping rule for the MH-sampler. As the clusters are defined by starting position, running MH-steps until there are no longer any cluster effects will minimise this variation. This has been borne out by calculations based on the simulations of Section 7. Therefore, one could run the MH-sampler until $\hat{V}_k(g)$ is suitably small.

The alternative to running long MCMC chains within each particle when there are new parameters in the model such as those introduced by the ‘shock’ at $t_k = 83$, is to expand the particle set by cloning each of the particles a number of times, each cloned particle having a fresh draw from the prior for each of the new parameter components. Upon observing the next batch of data, the expanded particle set could then be reduced down to a more manageable size. However, it is not clear *a priori* how many cloned copies of each particle to take and if the number of clones required exceeds the length of the parallel MCMC chains, then this does not represent a computational efficiency. Furthermore, this would not solve the problem in Scenario 2 where some parameters are not immediately identifiable.

A hybrid MH-kernel is introduced in Section 6. First, long-range, low-acceptance proposals are made, followed by short-range high-acceptance componentwise proposals. In many instances, this hybrid is replaced by a mixture distribution, a mixture of similar short and long-range moves. The adaptive proposal distributions of Fearnhead and Taylor (2013) might take this a step further, tuning the mixture probabilities so that the moves that have the largest expected jumps are proposed more often. This would be an attractive extension to this case, but through monitoring intra-class correlation it is clear that full block approximate Gibbs proposals maximise this expected jump size amongst the kernels we consider. However, we would still suggest moving at least a proportion of the particles according to random walk proposals, to guard against $\pi_k(\theta)$ being a degenerate approximation for $\pi_{k+1}(\theta)$.

A further problem of having step-changes in parameter values is the potential for a lack of identifiability. Scenario 2 provides two such examples. Firstly, we consider parameter η_2 , the dispersion in the immediate aftermath of $t_k = 83$. The prior for η_2 is a $\Gamma(0.01, 0.01)$ distribution chosen independently of η_1 (the dispersion preceding the shock). The sheer number of new parameters introduced in the aftermath of the shock ensures that the data on days 84, 85, 86 are (over-)fitted with very little error. The combination of this over-fitting and the unbounded nature of the prior close to zero pushes the initial posterior distributions for η_2 very close to zero. As data accumulate, the posterior mass gradually moves towards the value used to generate the data. This movement of posterior mass is difficult for sequential algorithms to track, particularly so because of the non-Gaussian nature of the prior, even on the log-scale. When performing real-time inference, therefore, the choice of a prior distribution more robust to this initial over-fitting may be preferred. Alternatively, a flat prior would need to be bounded to ensure that it can be sampled from. From a practical point of view, in the example of this paper, the choice of the $\Gamma(0.01, 0.01)$ distribution is meaningful as it attaches significant probability to the data being Poisson, rather than Negative Binomial, distributed.

The second example is the case of the background consultation parameters. The background rate of non-pandemic consultation is modelled using a log-linear spline, taking separate value for adults and children, with knots at $t_k = 84, 128, 176$, and 245 days. The value of the spline at these knots is given by

$$\mu + \alpha_t + \beta_a, \text{ s.t. } \sum_{t=1}^4 \alpha_t = \sum_{a=1}^2 \beta_a = 0,$$

with linear interpolation giving the value of the spline at the intervening points. This results in background consultation rates for days 84, 85, and 86 respectively of the form (neglecting the

age effects):

$$\begin{array}{c} \mu + \alpha_1 \\ \mu + 0.98\alpha_1 + 0.02\alpha_2 \\ \mu + 0.96\alpha_1 + 0.04\alpha_2 \\ \vdots \quad \quad \quad \vdots \end{array}$$

So, over this period there is very little identifiability of parameters μ and α_1 . This parameterisation, as shown in Figure 9, can induce convergence problems for MCMC but not for SMC. Jasra et al (2011) claim that, for their example, SMC may well be superior to MCMC and this is one case where this is certainly true. The population MCMC carried out in the rejuvenation stage achieves good coverage of the sample space, without the individual chains having to do likewise. To improve the MCMC mixing, this lack of identifiability would require a reparameterisation, which becomes unnecessary when using SMC.

8.2 Algorithmic Choice

Throughout, we have compared candidate MH-kernels via the KL-like statistics measuring the divergence between SMC posteriors from posteriors generated by the “gold-standard” MCMC. We have also constructed a reference distribution for the KL statistic to assess informally the significance of the observed divergences. This, however, rather presumes that the MCMC is the gold-standard. This superiority is, however, called into question by the better performance of the SMC algorithm particularly in the presence of the unidentifiability around shock times as discussed above.

From a computational efficiency point of view, the SMC algorithm, because of its highly parallel nature, is, at its worst, no slower than the full MCMC analysis. However, this may be an unfair comparison as the MCMC algorithm is based on “plain vanilla” random-walk Metropolis updates and could benefit from significant tuning itself. More sophisticated MCMC algorithms could be used, as exemplified in an epidemic context by Jewell et al. 2009. The use of differential geometric MCMC (Girolami & Calderhead, 2011) or advances in the parallelisation of MCMC (Banterle et al, 2015), for example, could assist with improving MCMC run times. On the other hand, as MCMC steps are the main computational overhead of the SMC algorithm, any development of the MCMC algorithm may lead to a similar improvement to the SMC algorithm also.

8.3 Data Availability

Up to now the discussion has centred on algorithmic development and the availability of all data sources in a timely manner has been assumed. Particularly crucial to the feasibility of real-time modelling is the role of the serology data. This is shown in Figure 10, where epidemic projections have been sequentially made using only noisy primary care consultation data in the absence of serological data. A clear and realistic picture of the epidemic is not available until the epidemic has almost entirely been observed. This poses some key questions: are serological samples going to be available in a timely manner, in sufficient quantity and quality, and in the right format? In reality, serological data can be slow to come online. A test has to be developed to identify the antibodies of a (probably) novel virus in blood sera; and there needs to be sufficient time to test samples and report results according to a protocol that ensures unbiased data collection and analysis.

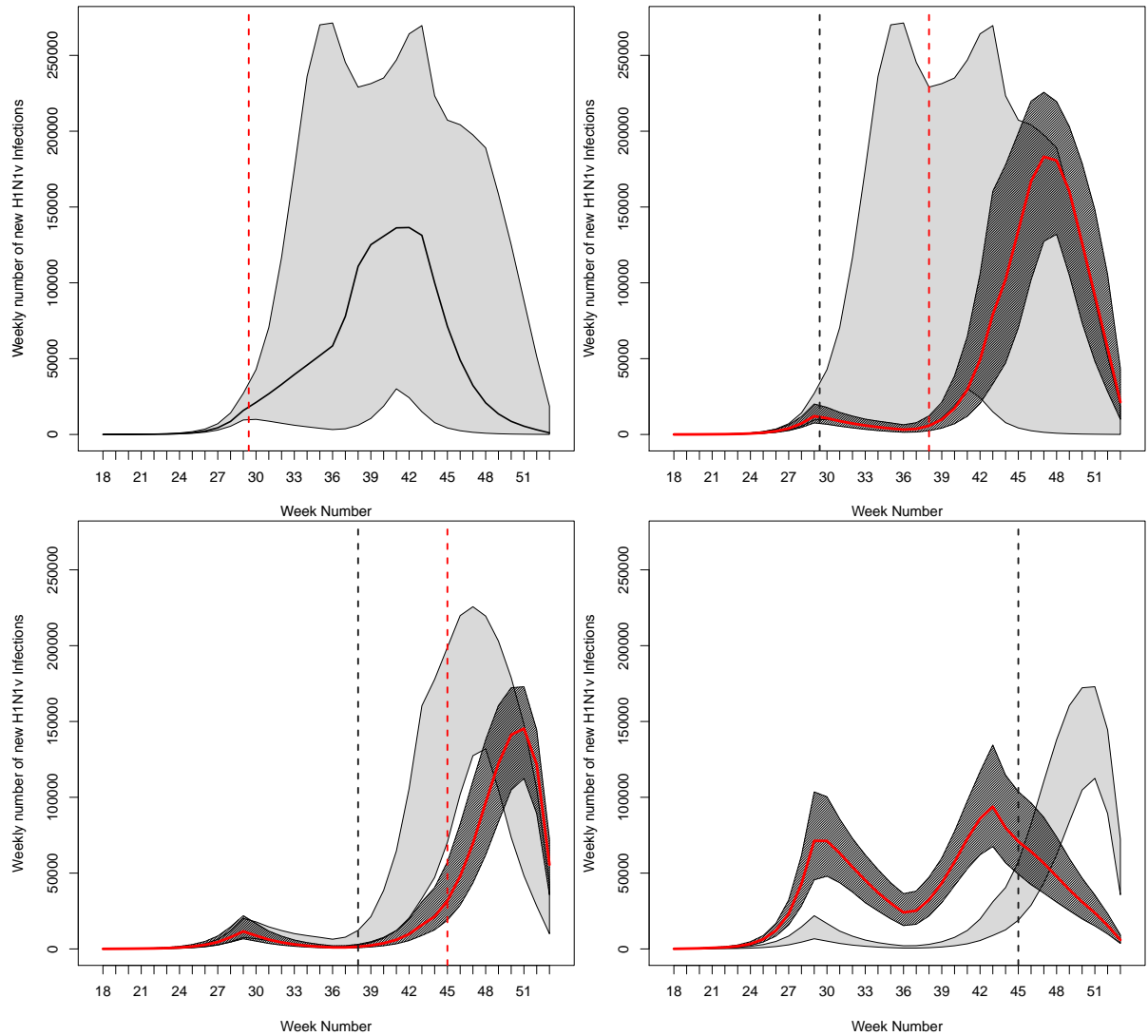


Figure 10: Sequential epidemic forecasts based on increasing amounts of data, not including serological data. The dark shaded areas represent a current forecast, with the light shaded areas the forecast at the time of the previous plot. Vertical red dashed lines indicate the current time, the black dashed line indicates the time of the previous prediction

From a computational point of view, under the assumption that all data become immediately available, each particle, in addition to its likelihood, weight and parameter value, stores a matrix representing the current state of the SEIR transmission model and a sub-history of values $\Delta_{t_k,a}$, long enough to evaluate Equation (4) at all current and future times. In the more realistic setting to accommodate the ‘slow’ serological data, particles will have to store the full historical values of $S_{t_k,a}$ in addition to the current state of the epidemic.

Finally, should external information that cannot be incorporated directly into the model become available at any time, it can easily be assimilated through appropriate adaptation of the prior distributions: particles would be reweighted according to the ratio of the new to old prior; and depending on the ESS, resampling and moving steps could follow. This provides a clear advantage of SMC over MCMC where the entire dataset would have to be re-analysed.

8.4 Stochasticity

The analyses in this paper have neglected the first fifty days of the epidemic, concentrating on a period when there is substantial transmission in the population and appropriate data are becoming available. As a result, a deterministic system can adequately describe the future evolution of the pandemic. Stochastic effects are significant and need to be incorporated into the model if monitoring is needed in the earlier stages. Amongst others, Nemeth et al. (2014) provide a prescription for particle learning in the presence of ‘shocks’ in such a setting. Alternatively, to improve the robustness of the inferences, the piecewise linear quantities describing population reporting behaviour $(p_{t_k,a}^P, B_{t_k,a})$ could be described by linked stochastic noise processes. This has the potential to reduce the sensitivity of estimates to the presence of changepoints that are not, for whatever reason, foreseeable.

8.5 Concluding Remarks

In answer to the question initially posed, we have provided a recipe for online tracking of an emergent epidemic using imperfect data from multiple sources. We have discussed many of the challenges to efficient inference, with particular focus on scenarios where the available information is rapidly evolving and is subject to sudden shocks. We have focused on an epidemic scenario likely to arise in the UK. Nevertheless, our approach addresses modelling concerns common globally (e.g. Shaman and Karspeck, 2012; Wu et al., 2010; Shubin et al., 2014; te Beest et al., 2015) and can form a flexible basis for real-time modelling strategies elsewhere. Real-time modelling is, however, more than just a computational problem. It does require the timely availability of relevant data, but also needs a sound understanding of any likely biases, and effective interaction with experts. In any country, only interdisciplinary collaboration between statisticians, epidemiologists and database managers can turn cutting edge methodology into a critical support tool for public health policy.

Acknowledgements

Paul Birrell was supported by the National Institute for Health Research (HTA Project: 11/46/03) the UK Medical Research Council (Unit Programme Numbers U105260566 and MC_UP_1302/3) and Public Health England.

References

- Ahrens, H. (1976) Multivariate variance-covariance components (MVCC) and generalized intraclass correlation coefficient (GICC). *Biometrical Journal*, **18**, 527–533. URL<http://dx.doi.org/10.1002/bimj.19760180703>.
- Banterle, M., Grazian, C., Lee, A. and Robert, C. P. (2015) Accelerating Metropolis-Hastings algorithms by Delayed Acceptance. *arXiv*, **1503.00996v2**, 27. URL<http://arxiv.org/abs/1503.00996v2>.
- Bettencourt, L. M. A. and Ribeiro, R. M. (2008) Real time Bayesian estimation of the epidemic potential of emerging infectious diseases. *PLoS ONE*, **3**, e2185.
- Birrell, P. J., Ketsetzis, G., Gay, N. G., Cooper, B. S., Presanis, A. M., Harris, R. J., Charlett, A., Zhang, X.-S., White, P., Pebody, R. G. and De Angelis, D. (2011) Bayesian modelling to unmask and predict the influenza A/H1N1pdm dynamics in London. *Proc. Natn. Acad. Sci. USA*, **108**, 18238–18243.
- Camacho, A., Kucharski, A., Aki-Sawyer, Y., White, M. A., Flasche, S., Baguelin, M., Pollington, T., Carney, J. R., Glover, R., Smout, E., Tiffany, A., Edmunds, J. J. and Funk, S. (2015) Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-time Modelling Study. *PLoS Currents*, **7**, Web. URL<http://dx.doi.org/10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2>.
- Cappé, O., Godsill, S. J. and Moulines, E. (2007) An overview of existing methods and recent advances in sequential Monte Carlo. *Proceedings of the IEEE*, **95**, 899–924.
- Carpenter, J., Clifford, P. and Fearnhead, P. (1999) Improved particle filter for non-linear problems. *IEE Proceedings - Radar, Sonar and Navigation*, **146**, 2+. URL<http://dx.doi.org/10.1049/ip-rsn:19990255>.
- Cauchemez, S., Boëlle, P.-Y., Thomas, G. and Valleron, A.-J. (2006) Estimating in real time the efficacy of measures to control emerging communicable diseases. *American Journal of Epidemiology*, **164**, 591–597.
- Chopin, N. (2002) A sequential particle filter method for static models. *Biometrika*, **89**, 539–552. URL<http://dx.doi.org/10.1093/biomet/89.3.539>.
- Del Moral, P., Doucet, A. and Jasra, A. (2006) Sequential Monte Carlo samplers. *J. R. Statist. Soc. B*, **68**, 411–436. URL<http://dx.doi.org/10.1111/j.1467-9868.2006.00553.x>.
- Donner, A. and Koval, J. J. (1980) The estimation of intraclass correlation in the analysis of family data. *Biometrics*, **36**, 19–25. URL<http://view.ncbi.nlm.nih.gov/pubmed/7370372>.
- Doucet, A. and Johansen, A. M. (2009) A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, **12**, 656–704.
- Dukic, V., Lopes, H. F. and Polson, N. G. (2012) Tracking epidemics with google flu trends data and a state-space seir model. *J. Am. Statist. Ass.*, **107**, 1410–1426. URL<http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2012.713876>.

- Dureau, J., Kalogeropoulos, K. and Baguelin, M. (2013) Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, **14**, 541–555. URL<http://dx.doi.org/10.1093/biostatistics/kxs052>.
- Farah, M., Birrell, P., Conti, S. and De Angelis, D. (2014) Bayesian emulation and calibration of a dynamic epidemic model for a/h1n1 influenza. *J. Am. Statist. Ass.*, **109**, 1398–1411. URL<http://www.ingentaconnect.com/content/tandf/uasa20/2014/00000109/00000508/art00009>.
- Fearnhead, P. (2002) MCMC, sufficient statistics and particle filters. *Journal of Computational and Graphical Statistics*, **11**, 848–862.
- Geyer, C. J. (1991) Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: The 23rd Symposium on the Interface*, 156–163. Interface Foundation of North America.
- Gilks, W. R. and Berzuini, C. (2001) Following a moving target - Monte Carlo inference for dynamic Bayesian models. *J. R. Statist. Soc. B*, **63**, 127–146. URL<http://dx.doi.org/10.1111/1467-9868.00280>.
- Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993) Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, **140**, 107–113. URL<http://ieeexplore.ieee.org/xpls/abs.all.jsp?arnumber=210672>.
- Jasra, A., Stephens, D. A., Doucet, A. and Tsagaris, T. (2011) Inference for Lévy-driven Stochastic Volatility Models via Adaptive Sequential Monte Carlo. *Scand. J. Statist.*, **38**, 1–22. URL<http://dx.doi.org/10.1111/j.1467-9469.2010.00723.x>.
- Jasra, A., Stephens, D. A. and Holmes, C. C. (2007) On population-based simulation for static inference. *Statistics and Computing*, **17**, 263–279. URL<http://dx.doi.org/10.1007/s11222-007-9028-9>.
- Jewell, C. P., Kypraios, T., Christley, R. M. and Roberts, G. O. (2009) A novel approach to real-time risk prediction for emerging infectious diseases: a case study in Avian Influenza H5N1. *Preventive veterinary medicine*, **91**, 19–28. URL<http://dx.doi.org/10.1016/j.prevetmed.2009.05.019>.
- Kantas, N., Beskos, A. and Jasra, A. (2014) Sequential monte carlo methods for high-dimensional inverse problems: A case study for the navier–stokes equations. *SIAM/ASA Journal on Uncertainty Quantification*, **2**, 464–489. URL<http://dx.doi.org/10.1137/130930364>.
- Konishi, S., Khatri, C. and Rao, C. R. (1991) Inferences on multivariate measures of interclass and intraclass correlations in familial data. *J. R. Statist. Soc. B*, 649–659.
- Liang, F. and Wong, W. H. (2001) Real-Parameter Evolutionary Monte Carlo With Applications to Bayesian Mixture Models. *J. Am. Statist. Ass.*, **96**, 653–666. URL<http://dx.doi.org/10.1198/016214501753168325>.
- Liu, J. S. and Chen, R. (1995) Blind Deconvolution via Sequential Imputations. *J. Am. Statist. Ass.*, **90**, 567–576. URL<http://dx.doi.org/10.2307/2291068>.
- (1998) Sequential Monte Carlo Methods for Dynamic Systems. *J. Am. Statist. Ass.*, **93**, 1032–1044. URL<http://dx.doi.org/10.2307/2669847>.

- Martin, J. S. (2012) *Some New Results in Sequential Monte Carlo*. Ph.D. thesis, Department of Mathematics, Imperial College.
- Meester, R., de Koning, J., de Jong, M. C. and Diekmann, O. (2002) Modeling and real-time prediction of classical swine fever epidemics. *Biometrics*, **58**, 178–184. URL<http://view.ncbi.nlm.nih.gov/pubmed/11892689>.
- Neal, R. M. (1996) Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, **6**, 353–366. URL<http://dx.doi.org/10.1007/bf00143556>.
- Nemeth, C., Fearnhead, P. and Mihaylova, L. (2014) Sequential Monte Carlo Methods for State and Parameter Estimation in Abruptly Changing Environments. *IEEE Transactions on Signal Processing*, **62**, 1245–1255. URL<http://dx.doi.org/10.1109/tsp.2013.2296278>.
- Ong, J. B. S., Chen, M. I.-C., Cook, A. R., Chyi, H., Lee, V. J., Pin, R. T., Ananth, P. and Gan, L. (2010) Real-time epidemic monitoring and forecasting of H1N1-2009 using influenza-like illness from general practice and family doctor clinics in Singapore. *PloS one*, **5**, e10036. URL<http://dx.doi.org/10.1371/journal.pone.0010036>.
- Roberts, G. O. and Rosenthal, J. S. (2001) Optimal Scaling for Various Metropolis-Hastings Algorithms. *Statistical Science*, **16**, 351–367. URL<http://dx.doi.org/10.2307/3182776>.
- Scientific Pandemic Influenza Advisory Committee (SPI): Subgroup on Modelling (2011) Modelling Summary. URLhttp://webarchive.nationalarchives.gov.uk/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/@dh/@ab/documents/digitalasset/dh_127275.pdf. SPI-M-O Committee document (Accessed 4 February, 2016).
- Shaman, J. and Karspeck, A. (2012) Forecasting seasonal outbreaks of influenza. *Proc. Natn. Acad. Sci. USA*, **109**, 20425–20430. URL<http://www.pnas.org/content/early/2012/11/21/1208772109.abstract>.
- Sherlock, C., Fearnhead, P. and Roberts, G. O. (2010) The random walk metropolis: Linking theory and practice through a case study. *Statistical Science*, **25**, 172–190.
- Shubin, M., Virtanen, M., Toikkanen, S., Lyytikäinen, O. and Auranen, K. (2014) Estimating the burden of A(H1N1)pdm09 influenza in Finland during two seasons. *Epidemiology and infection*, **142**, 964–974. URL<http://dx.doi.org/10.1017/s0950268813002537>.
- Skvortsov, A. and Ristic, B. (2012) Monitoring and prediction of an epidemic outbreak using syndromic observations. *Mathematical Biosciences*, **240**, 12–19. URL<http://dx.doi.org/10.1016/j.mbs.2012.05.010>.
- Sokal, R. R. and Rohlf, F. (1981) *Biometry* (2nd edn). New York: WH Feeman and Company, **668**, .
- te Beest, D. E., Birrell, P. J., Wallinga, J., De Angelis, D. and van Boven, M. (2015) Joint modelling of serological and hospitalization data reveals that high levels of pre-existing immunity and school holidays shaped the influenza A pandemic of 2009 in the Netherlands. *Journal of the Royal Society, Interface*, **12**, 20141244+. URL<http://dx.doi.org/10.1098/rsif.2014.1244>.

- Wallinga, J. and Teunis, P. (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology*, **160**, 509–516.
- Wu, J. T., Cowling, B. J., Lau, E. H. Y., Ip, D. K. M., Ho, L.-M., Tsang, T., Chuang, S.-K., Leung, P.-Y., Lo, S.-V., Lio, S.-H. and Riley, S. (2010) School Closure and Mitigation of Pandemic (H1N1) 2009, Hong Kong. *Emerging Infectious Diseases*, **16**, 538–541.